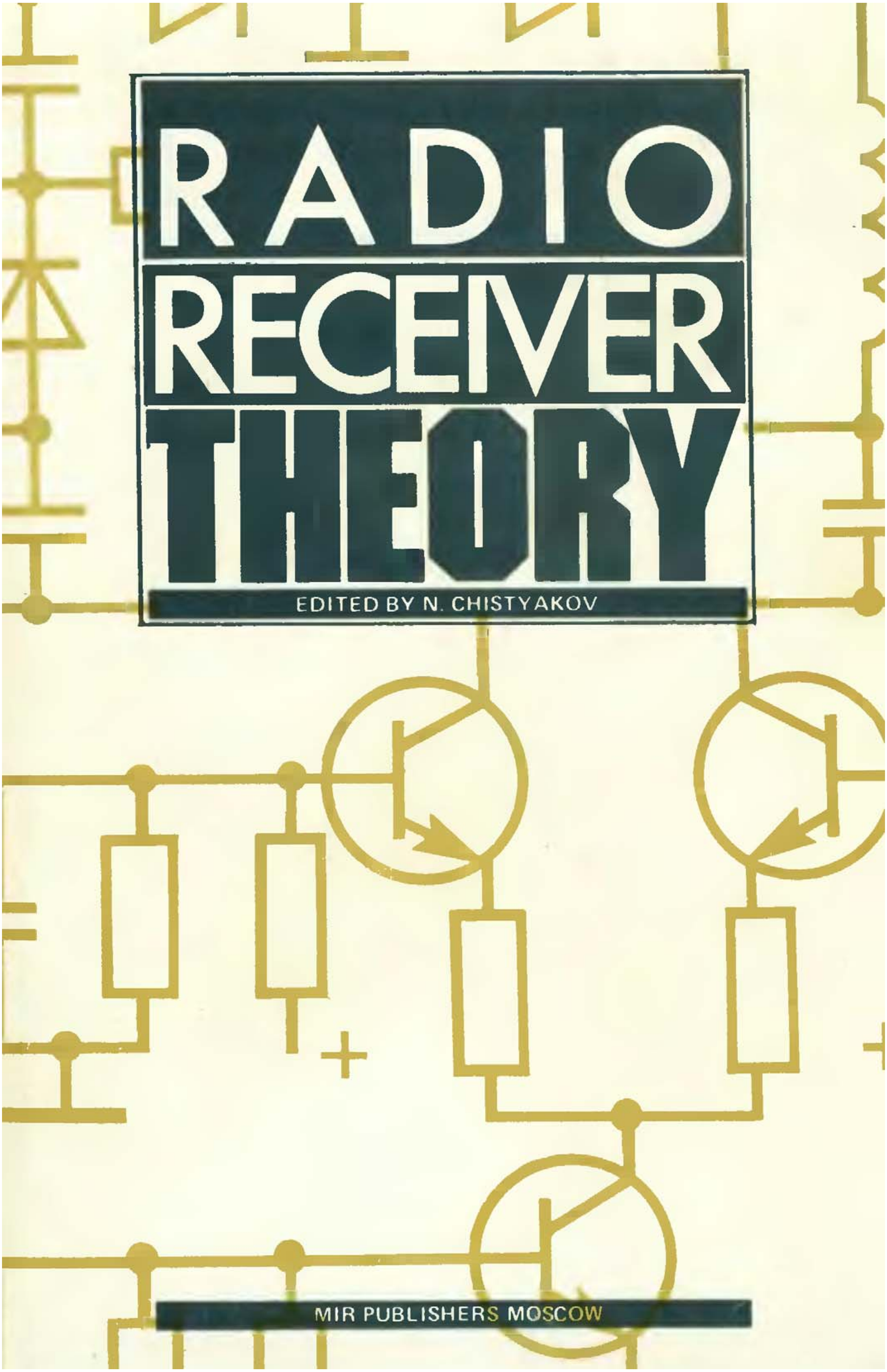


RADIO RECEIVER THEORY

EDITED BY N. CHISTYAKOV

MIR PUBLISHERS MOSCOW

РАДИОПРИЕМНЫЕ УСТРОЙСТВА

Н. Н. Буга, А. И. Фалько, Н. И. Чистяков

Под общей редакцией Н. И. Чистякова

Москва «Радио и связь»

N. BUGA, A. FALKO, N. CHISTYAKOV

RADIO RECEIVER THEORY

Edited by N.I. Chistyakov

Mir Publishers Moscow

First published 1990
Revised from the 1986 Russian edition

Translated from the Russian
by Boris V. Kuznetsov

На английском языке

ISBN 5-03-001321-0

Contents

Preface

This book is a general course on radio receivers as taught in the Soviet Union to all students majoring in radio communications and broadcasting. Those who plan to specialize in radio engineering take an additional course concerned with radio receiver design and operation and based on the present course. For this reason, this book is primarily concerned with the principles underlying radio reception and radio receiver theory and does not go into circuit and component details.

The authors have widely drawn on related subjects, such as the theory of linear and nonlinear circuits and electron devices, the theory of telecommunications, etc., presumably already taken by the reader as part of his studies.

In a way, the book reflects the experience gained in the field by the radio receiver departments at the Leningrad, Moscow and Novosibirsk Telecommunication Institutes.

The authors have contributed the material as follows.

N. N. Buga: Introduction, Chapter 1 (except Sec. 1.7), Chapter 7, Sec. 9.5, and Sec. 10.9.

N.I. Chistyakov: Secs. 2.4 and 3.13, Chapters 4, 6, and 8 (except Secs. 8.4 and 8.11), Chapter 9 (except Secs. 9.5 through 9.7), Chapter 10 (except Sec. 10.9), and Conclusion.

A.I. Falko: Sec. 1.7, Chapters 2 and 3 (except Secs. 2.4 and 3.13), Chapter 5, and Secs. 8.4, 8.11, 9.6 and 9.7.

The Authors

Introduction

Any telecommunication system has as one of its key elements what we will simply call a receiver. For the purpose of our discussion, we will also use this term to describe a combination of a receiving antenna, a receiver proper, and an end device. The antenna picks up electromagnetic energy and converts it to a radio-frequency (r.f.) voltage. The receiver processes the incoming energy spectrum to extract valid signals, amplifies them at the expense of a local power supply, processes them so as to minimize the noise or other interference present in the input wave, and detects the r.f. signals to shape a wave which is a faithful replica of the transmitted message. In the end device, the signals thus extracted are utilized to produce a desired effect which may be audio (by a speaker or a pair of headphones), visual (by a TV picture tube or kinescope), mechanical (by a radio telegraph set), etc. The end device may be an integral part of the receiver or a stand-alone unit.

In this book, we will look into the physical foundations of signal reception in the presence of noise, the principles underlying the structure of various receivers and their key functional elements, their theory, and parameter calculation. Presumably, the reader has acquired a working knowledge of the mathematical models and properties applicable to signal and noise, the methods used in the analysis of linear and nonlinear circuits and communication channels having constant and varying parameters, the theory and properties of antenna-feeder systems, radio wave propagation, the physical principles lying at the basis of signal generation, amplification, modulation and detection, frequency conversion, amplitude limiting, the basic theory of electromagnetic compatibility in relation to telecommunications, integrated-circuit (IC) technology, and use of microprocessors.

The structure and key functions of a receiver depend on the conditions of signal reception. What goes from an antenna to the receiver input is a mixture of a wanted signal, and noise. Noise may come from other radio stations, industrial sources, electromagnetic processes taking place in the atmosphere and outer space, and thermal radiation of the Earth, and may be many times as strong as the wanted signal in the antenna. The signals themselves may suffer amplitude and phase distortion in the wake of changes in the conditions of radio wave propagation. A receiver should be able to separate wanted signals from noise on the basis of features inherent in the signals. This property is called *selectivity*.

There are several kinds of selectivity, as follows.

Frequency selectivity. It is obtained with the aid of frequency-selective circuits. Since frequency allocation among the various services and among the systems within each service is subject to national and international regulations, any receiver is expected to possess the required level of frequency selectivity.

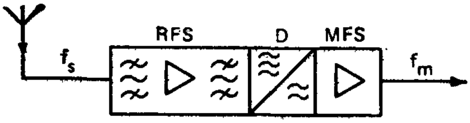


Fig. I.1

Spatial selectivity. It is obtained with the aid of directional antennas.

Polarization selectivity. It is obtained with the aid of antennas which pick up waves having a particular type of polarization.

Amplitudes selectivity. It is primarily used in the reception of pulse signals by a special type of threshold circuits known as amplitude selectors.

Time selectivity. It is obtained by enabling the receiver for the duration of a time slot during which a wanted signal is expected to arrive.

Signal-shape selectivity. Among other things, it may be based on code structure of the signal.

On the basis of the foregoing, we may depict a receiver as shown in Fig. I.1. Here, a detector D separates the receiver into two sections: a radio-frequency section, RFS, and a series-connected modulation frequency section, MFS. The RFS is responsible for frequency selection and amplification; it may also perform frequency conversion, amplitude and time selection, and signal conditioning so as to minimize the effect of noise and interference. The MFS does post-detector signal processing, namely: amplification and, say, integration so as to minimize the effect of noise, decoding, and channel separation as in multichannel systems. In the latter case, as can be seen from Fig. I.2, the receiver has a common section made up by an r.f. section RFS, a detector D, and a modulation-frequency section MFS which form between them a group modulation-frequency amplifier, GMFA. It is followed by a channel separator CS, demodulators DM_1 through DM_n, and modulation-frequency amplifiers MFA_1 through MFA_n.

Receivers may be classed in more than one way. What follows is the classification adopted in the Soviet Union.

By function, they are classed into professional and broadcasting, or home, receivers. Professional receivers include those used for point-to-point, radio-relay, zone, local and other forms of radio communications, radio astronomy, radar, radio navigation, etc. Broadcasting receivers serve to receive radio and TV broadcasts and are the most numerous ones.

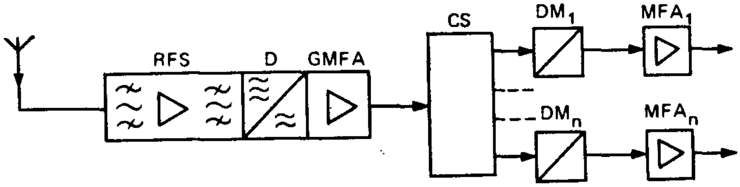


Fig. I.2

By type of modulation, receivers are classed into amplitude-modulated (AM), frequency-modulated (FM), phase-modulated (PM), etc.

By frequency or wavelength, they are classed according to the nomenclature of frequency bands given for reference in the table that tollows.

Nomenclature for Frequency Bands

Frequency range	Abbreviation	Meaning	Metric designation by wavelength
30-300 Hz	ELF	Extremely low frequency	Megametric waves
300-3000 Hz	VF	Voice frequency	—
3-30 kHz	VLF	Very low frequency	Myriametric waves
30-300 kHz	LF	Low frequency	Kilometric waves
300-3000 kHz	MF	Medium frequency	Hectometric waves
3-30 MHz	HF	High frequency	Decametric waves
30-300 MHz	VHF	Very high frequency	Metric waves
300-3000 MHz	UHF	Ultra-high frequency	Decimetric waves
3-30 GHz	SHF	Super-high frequency	Centimetric waves
30-300 GHz	EHF	Extremely high frequency	Millimetric waves
300-3000 GHz or 3 terahertz	—	—	Decimillimetric waves

Accordingly, there are LW or LF, MW or MF, SW or HF, VHF, UHF, SHF, etc. receivers. Some are built as all-wave units which can operate in several frequency bands.

By kind of information they handle, receivers are classed into radio-telephone, radio-telegraph, facsimile, TV, and some other types. Some receivers can handle several kinds of information.

There is a further classification of receivers according to the services in which they are used. These services are defined as follows: *fixed* in which radio communication is between specified fixed points (examples are point-to-point HF circuits and microwave links), and *mobile* in which radio communication is between stations intended to be used while in motion or during halts at unspecified points or between such stations and fixed stations. In the latter case, a further subdivision is into land mobile (exemplified by automobile receivers), maritime mobile, aeronautical mobile, and space mobile.

Finally, there may be manually (that is, locally) and remotely operated, attended and unattended receivers.

For their operation receivers may draw power from an a.c. supply line (mains) or a storage battery, or both (as in the case of a.c./d.c. receivers).

Advances in receiver design and performance have inseparably been linked to progress in electronic technology and components. Since the early '50s receivers have become largely transistorized, the only exception being TV and radar receivers which are still using vacuum and cathode-ray tubes. Strong impetus to advances in radio reception has come from integrated-circuit (IC) electronics, digital signal processing, computers, and the use of ever higher frequencies in the microwave band. The advent of medium- and large-scale integration, followed by very large scale (VLS) and extremely large scale (ELS) integration, has done much to improve the performance and ergonomics of receivers as well. Microprocessors offers a means with which to automate receiver operation, implement efficacious signal processing and noise situation analysis, and to use the results for adaptive receiver control. All of the above factors go a long way towards extending receiver capabilities, simplifying their manufacture, and making them simpler to operate and maintain.

The '60s saw the arrival of space communications accompanied by improvements in microwave receivers and a greater reliance on low-noise maser and parametric amplifiers. The recent decades have been characterized by a further drive into the millimetric (EHF) and optical wavelength bands.

A crucial problem today is to ensure electromagnetic compatibility of telecommunication systems. This is especially urgent for radio communications and broadcasting in the HF band. The progressively growing use of the SHF band inevitably poses similar problems for SHF operators as well.

Electromagnetic compatibility implies noise control at both source and receiver. That is why special emphasis in receiver theory is placed on the evaluation of receiver noise immunity, receiver noise suppression, receiver noise compensation, optimal signal processing, and some other related issues.

Chapter One

Receiver Performance Characteristics and Structure

1.1. The Receiver as a Part of a Complex System

Any receiver has properties common to the subsystems of a complex system. Notably, it interacts with the other elements of the system of which it is a part, and with the surroundings; it has a hierarchical structure; and it operates in random conditions.

Apart from a receiver, a typical telecommunication system would generally include a transmitter at the sending end, an antenna switch, an operator's remote-control console, a control computer, monitors and recorders, measuring instruments, and power supplies. The interaction of the receiver with other system elements and with the medium through which radio signals travel towards the receiving end involves the reception of valid signals, the interfering signals produced by the system elements (in-system interference) and radiations from external sources, that is, those which are not part of a given system (unintentional and intentional interference), the action of the receiver's information (signal) outputs on the operator and the automatic control facilities, and the reaction of these facilities and of the operator to the signal and power-supply inputs of the receiver.

A receiver is said to have a hierarchical structure because its units effect control over some and are at the same time controlled by other units; that is, the receiver units are arranged and operate according to rank or order. This feature stands out with especial clarity in receivers having various automatic control circuits.

A receiver is said to behave in a random manner, or stochastically, because wanted signals are always picked up with no or little prior knowledge about noise or interference. Therefore, signal identification is a probabilistic procedure, and the final decision-making may involve a series of partial results yielded by signal processing (as in the receivers of adaptive telecommunication systems using noise-immune signal coding).

As a subsystem of a complex system, a receiver can be described by a set of external and internal variables. The external variables characterize the interaction of the receiver with other elements of the system and with the surroundings, while the internal variables characterize the configuration, functioning, and the dynamic and structural relations between the receiver units. The external variables describe the performance of a receiver from the viewpoint of the customer, while the internal variables do so from the viewpoint of its designer.

The external variables include the frequency band in which a given receiver operates, the form of the received signals, sensitivity, susceptibility to noise and interference, the level of own spurious radiation, selectivity, noise immunity, fidelity, frequency-setting accuracy, output-signal power and shape, the variables related to the design and operation of the receiver (performance stability ergonomics, reliability, maintainability, power requirements, mobility, size, weight, cost, etc.). The internal variables include the number and limits of frequency bands, the dynamic range, bandwidth, gain, etc.

Both the external and the internal variables are subject to constraints which are taken into account in optimizing receiver synthesis.

If, under given service conditions and with all other variables held constant, an improvement in the performance of receiver S calls for a decrease or an increase in some external variable k, the latter may be taken as directly bearing on the quality of the receiver.

There may be scalar synthesis and vector synthesis. In scalar synthesis, receiver S is characterized by an only number, the variable k (S). This may be the cost C: assuming that all the other variables remain unchanged, receiver S_1 will show a better performance than receiver S_2 if $C_1 < C_2$.

In vector synthesis, one takes into account the fact that any receiver is a complex device and that its performance is affected by several interrelated variables. Such a receiver should be optimized on the basis of a set of quality variables that form between them a quality vector, $\mathbf{K} = | k_1, \ldots, k_m |$. This means that in vector synthesis the quality of a receiver is characterized by an ordered set of m numbers rather than by one number k as in scalar synthesis. Whereas in scalar synthesis the quality variable k offers a basis for an unambiguous comparison to be made between receivers S_1 and S_2, no comparison may prove feasible on the basis of the quality vector $\mathbf{K}$ in some cases unless resort is made to further preference criteria because any direct comparison may turn out inapplicable to $\mathbf{K}_1$ (S_1) and $\mathbf{K}_2$ (S_2).

This point can best be illustrated by an example. In comparing S_1 and S_2 it is presumed that S_1 performs better than S_2 if each of the quality variables, k_i (S_1), $i = 1, \ldots, m$, of receiver S_1 is not worse (not greater) than each of the quality variables of receiver S_2. Let now the two receivers be characterized by the quality vectors $\mathbf{K}_1 = | 3, 15 |$ and $\mathbf{K}_2 = | 3, 20 |$. Then it may be argued that receiver S_1 is unconditionally better than receiver S_2. Quite aptly, this condition may be referred to as an unconditional preference criterion. If, on the other hand, $\mathbf{K}_1 = | 3, 15 |$, and $\mathbf{K}_2 = | 5, 11 |$, the two receivers cannot be compared vectorially, and one has to invoke a conditional preference criterion, proceeding from the purposes that are to be served by the receivers. Conditional preference criteria may be widely different for different types of receivers. For example, this may be quality vector $K = | k_1, \ldots, k_m |$ for which the weighted sum of the individual quality variables, $a_1k_1 + \ldots + a_mk_m$, is minimal.

In summary, if design optimization finally leads to a 10% reduction in the cost of a broadcast receiver, the ultimate saving in the overall cost would be impressive, indeed, considering the scale on which this type of receiver is manufactured and sold.

1.2. Block Diagrams of Receivers

A receiver will be represented by a different block diagram, depending on whether it is a *tuned radio frequency* (TRF) type or a *superheterodyne* (or *superhet*) type. These two types differ in the arrangement of the r.f. section.

In a TRF receiver, as can be seen from Fig. 1.1, the r.f. section, RFS, contains an input circuit, IC, and a radio-frequency amplifier, RFA, to boost the signal picked up by the antenna. In this case, all resonant circuits are tuned to the signal, or, radio frequency. The input circuit provides frequency pre-selection ahead of the RFA, and the RFA supplies the bulk of frequency selection and pre-detector signal amplification. The resonant networks within the input circuit and the r.f. amplifier can be tuned to any desired frequency within the operating frequency range. Since there is a need for high selectivity and high gain (the voltage gain of the r.f. amplifier may be as high as 10^6 or 10^7), several amplifying stages and tuned circuits will be usually provided. Because of the difficulties in providing gang tuning and proper tracking, the number of tuned circuits is seldom greater than three or four. In the circumstances, the gain at the radio or signal frequency, f_s, may prove unstable and the selectivity insufficient because the bandwidth B of a tuned circuit is

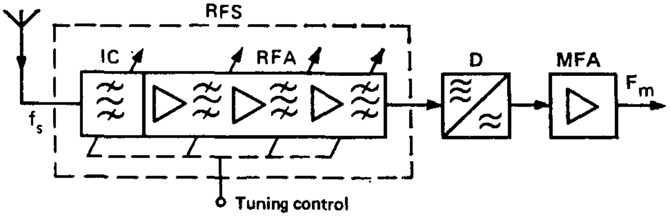


Fig. 1.1

connected to its resonance frequency, $f_0 = f_s$, as

$$B = f_0/Q$$

where Q is the quality factor (a figure of merit) of the tuned circuit. With variations in the frequency to which a given resonant circuit is tuned, the selectivity and the gain also vary (as f_s increases, B also increases and, as a consequence, the selectivity decreases).

As a way of reducing the number of amplifying stages and simplifying the receiver design, it was customary in the past to use regenerative and superregenerative amplifiers in the r.f. section of TRF receivers.

In TRF receivers using a regenerative amplifier, positive or regenerative feedback introduces in the tuned circuit a negative resistance which makes up for losses and serves thus to increase the gain. Unfortunately, such receivers are of low stability because they operate under conditions close to those which are likely to cause the amplifier to oscillate. The oscillations thus generated may break through to the antenna and be radiated, thus interfering with the operation of other receivers, which is highly undesirable from the viewpoint of electromagnetic compatibility. Regenerative amplifiers in which negative resistance is derived by quite different principles are used in state-of-the-art microwave receivers. They are built around nonreciprocal devices called circulators, which suppress parasitic radiations.

In a superregenerative receiver, positive feedback in the RFA is periodically varied at an auxiliary frequency which is substantially higher than the modulation frequency. Because of this, the coupled-in resistance turns negative during a part of a cycle, and oscillations are induced in the tuned circuit. Their amplitude is 10^4 or more times the signal amplitude. Their strength is proportional to that of the incoming signal, which means that the oscillations thus generated are in effect the amplified valid signals. Unfortunately, superregenerative receivers, similarly to the regenerative type, face the problem of parasitic oscillations, because of which they fail to meet the requirements of electromagnetic compatibility. This limits their use, unless measures have been taken to prevent parasitic oscillations from reaching the antenna.

The superhet receiver is the most commonly used type. In it, as can be seen from Fig. 1.2*a*, the signal frequency f_s is converted by a frequency converter (also called a frequency changer), FC, which consists of a local oscillator, LO, and a mixer, Mxr, to what is called the intermediate frequency, f_i. It is at the intermediate frequency (usually abbreviated as the i.f.) that the major portion of amplification and frequency selection is effected. For this reason, the pre-detector section consists of two parts. One is the r.f. section. RFS, containing the input circuit, IC, and an r.f. amplifier, RFA, and the other is the i.f. section, IFS, which contains the frequency converter, FC, and an i.f. amplifier, IFA.

The mixer only changes the signal frequency and does not affect the waveform of the modulation function, which is another way of saying that it acts as a linear parametric network as regards the incoming signal.

When the inputs of the mixer are fed the signal frequency f_s and the local-oscillator frequency f_{LO}, its output delivers what are called intermodulation frequencies $mf_s + nf_{LO}$, where m and n may be equal to ± 1, ± 2, etc. The resonant circuit provided at the mixer output may be tuned to either the difference (difference conversion) or the sum (sum conversion) of f_s and f_{LO} (sometimes, $2f_{LO}$, $3f_{LO}$, etc.). With difference conversion

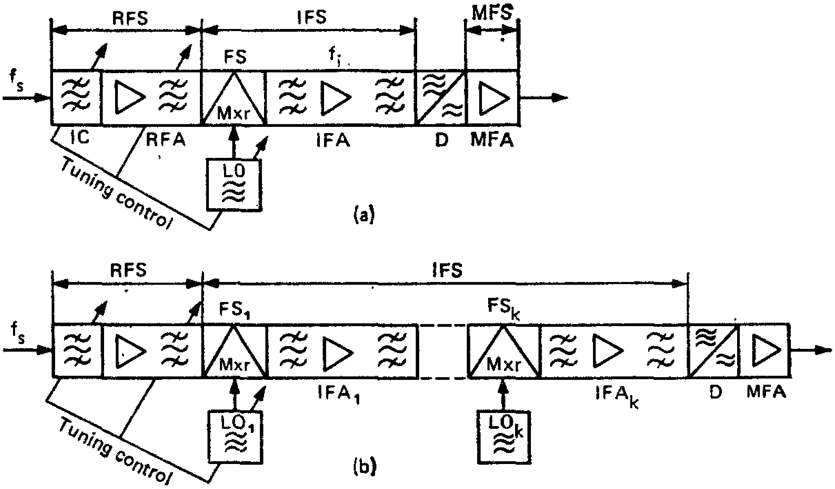


Fig. 1.2

$$f_i = f_s - f_{LO}$$

or

$$f_i = f_{LO} - f_s$$

When $m = 1$ and $n = -1$, the local oscillator is set to a frequency which lies above f_s and it is said to be "down-tuned". When $m = -1$ and $n = 1$, the local oscillator is set to a frequency which lies below f_s, and it is said to be "up-tuned". In either case, f_{LO} can be chosen such that the intermediate frequency f_i is below the operating frequency range, $f_i < f_{s,\min}$.

Frequency conversion with upward translation of the signal spectrum, in which case $f_i > f_{s,\max}$, is feasible with both difference and sum conversion. When the i.f. is higher than the signal, the receiver is referred to as an *infradyne receiver*. The high i.f. has then to be down-translated in another converter, so infradyne receivers use multiple frequency conversion.

If the signal frequency f_s falls within some fixed frequency band, the intermediate frequency f_i can then be maintained constant by varying f_{LO}. This is achieved by means of what is known as gang tuning of the input circuit, the r.f. amplifier and the local oscillator with a single mechanical control.

The fact that the varying signal frequency f_s is converted in a superhet receiver to a constant intermediate frequency f_i provides for this type of receiver a number of advantages, namely:

— the resonant circuits in the i.f. section need not be re-tuned as the signal frequency varies, and this simplifies their construction; since the gain is held at a constant value, the overall gain of the receiver only slightly depends on the frequency to which it is tuned;

— down-conversion (that is, frequency conversion with the signal spectrum translated downwards) mitigates the effect of parasitic capacitive and inductive feedback; this serves to enhance gain without impairing stability;

— operation at a down-converted f_i offers a chance to reduce bandwidth and enhance selectivity without untoward complication in resonant-circuit construction.

Proceeding from the foregoing, the following differences can be noted between the resonant (tuned) circuits used in the r.f. and i.f. sections.

(1) The r.f. tuned circuits have in most cases a relatively broad bandwidth so that it may encompass adjacent-channel frequencies in addition to the radio signal spectra. These circuits effect preselection, therefore the receiver's section containing the input circuit and the r.f. amplifier is often referred to as a preselector.

(2) The i.f. tuned circuits have a bandwidth comparable with that of the signal spectrum and suppress unwanted interfering signals falling outside that spectrum.

Frequency conversion endows superheterodyne reception with a number of special features, such as the appearance of what is differently called as spurious outputs, spurious responses or spurious reception channels, the effect of local-oscillator frequency on tuning, and the likelihood of local-oscillator signals to be radiated by the receiving antenna.

The receiver bandwidth encompassing the signal spectrum forms the wanted reception channel. The frequency bands which are adjacent to the main channel and which may be occupied by undesired spectra form adjacent channels.

The intermediate frequency can be derived not only by the conversion of the signal frequency as defined above, but also as an interfering frequency f_{int} beats with the local-ocillator frequency, according to the equation

$$mf_{\mathrm{int}} + nf_{\mathrm{LO}} = f_{\mathrm{i}}$$

where m and n are equal to 0, ± 1, ± 2, etc.

As the interference falls within the bandwidth of the i.f. section secondary to frequency conversion by the above equation, it is superimposed on the incoming signal and distorts it. The frequency bands within which false signals are likely to arise constitute what we have called spurious responses or channels. The worst offenders among them are image interference or, simply, the image, and interference at the i.f. If, for example, the desired frequency conversion is such that

$$f_{\mathrm{i}} = f_{\mathrm{LO}} - f_{\mathrm{s}} \text{ for } m = -1,\ n = 1$$

there is a chance for a spurious response to take place by the equation

$$f_{\mathrm{i}} = f_{\mathrm{int}} - f_{\mathrm{LO}}$$

if $m = 1$ and $n = -1$. This has been referred to as the image, defined as $f_{\mathrm{int}} = f_{\mathrm{im}}$. In this case, the valid signal f_{s} lies below f_{LO} by f_{i}, and the image f_{im} will then lie above f_{LO} by f_{i}, thus showing mirror-like symmetry about the local-oscillator (or heterodyne) frequency. If, on the other hand, $f_{\mathrm{s}} = f_{\mathrm{LO}} + f_{\mathrm{i}}$, then $f_{\mathrm{im}} = f_{\mathrm{LO}} - f_{\mathrm{i}}$.

If the interfering frequency is the same as the intermediate frequency, f_{i}, to which the i.f. section is tuned and if this interference has not been suppressed in the preselector, it will, similarly to the image, reach the i.f. section and will be amplified there. This case ($m = 1$ and $n = 0$) has been referred to as i.f. interference or i.f. spurious response.

There are also spurious responses associated with the harmonics of the local-oscillator frequency ($2f_{\mathrm{LO}}$, $3f_{\mathrm{LO}}$, etc.).

To prevent spurious responses from reaching the i.f. section, they must be suppressed by selective networks ahead of the frequency converter.

The requirement for high adjacent-channel and image selectivity often drives one to use several (two or even three) frequency conversions in a receiver. Quite logically, this involves the use of several frequency converters, as shown in Fig. 1.2*b*. Each converter produces a converted or intermediate frequency of its own, which is then translated to the final intermediate frequency in the last frequency converter. The final i.f. section usually provides the required adjacent-channel selectivity and the major portion of amplification. Because spurious responses arise with each frequency conversion, there is a need to suppress the interference associated with them by selective networks ahead of the respective frequency converter.

There is a frequency stabilization system applied to all the frequency converters. It is a necessary adjunct because variations in the local-oscillator frequency would have caused a proportionate deviation of the intermediate frequency from its desired or nominal value, $f_{i,n}$. Suppose that the nominal local-oscillator frequency $f_{LO,n}$ has changed by Δf_{LO}. Then.

$$f_i = f_{LO} - f_s = (f_{LO,n} + \Delta f_{LO}) - f_s = f_{i,n} + \Delta f_{LO}$$

If the local oscillator operates at a relatively high frequency, the deviation of f_i from its nominal value $f_{i,n}$ to which the selective networks in the i.f. section are tuned might be considerable indeed, entailing an impairment in the gain of the section and, as a corollary, an impairment in receiver sensitivity. The deviation of the signal spectrum relative to the receiver bandwidth might either lead to the corruption of the received signal or cause an interfering signal to fall within the passband of the i.f. amplifier instead of the valid signal.

1.3. The Operating Frequency Range

The term 'operating frequency range' refers to the range of frequencies within which a receiver is tunable. With continuous tuning, the operating frequency range is defined by what are known as the upper and lower edge frequencies, $f_{0,max}$ and $f_{0,min}$. The relative width of the operating frequency range is stated in terms of the maximum-to-minimum frequency ratio defined as

$$k_r = f_{0,max}/f_{0,min}$$

To ensure a high value of k_r, simple tuning, and high-quality reception, it is usual to break down the overall frequency range into several bands such that their individual maximum-to-minimum frequency ratio is

$$k_{b,i} = (f_{b,max}/f_{b,min})_i$$

The value of k_b is above all limited by the structural capabilities of tuning capacitors for which C_{max} is usually about 25 to 50 times C_{min}. Then

$$k_b = (C_{max}/C_{min})^{1/2} \approx 5 \text{ to } 7$$

Considering that the stray wiring capacitance is added to C_{min} and in view of the expected use of the receiver, it is usual to take $k_b \leqslant 2$ or 3. As a rule, k_b is decreased with increasing operating frequencies and quality of reception.

The operating frequency range may be divided into N bands so as to obtain either equal maximum-to-minimum frequency ratios, $k_{b,i}$, or the same frequency span for all the bands. In the former case, $k_{b,i}$ = const, and $k_r = k_b^N$. Hence,

$$k_b = (f_{0,\max}/f_{0,\min})^{1/N}$$

The number of bands is given by

$$N = \log_{10}k_r/\log_{10}k_b$$

It is an easy matter to see that the frequency span for the Nth band is

$$\Delta f_{b,N} = \Delta f_{b1}k_b^{N-1}$$

where Δf_{b1} is the span of the first band. Thus, the band span increases with increasing band number, N.

In the latter case, $\Delta f_{b,i}$ = const, but the maximum-to-minimum frequency ratios $k_{b,i}$ are all different:

$$k_{b,N} = 1 + \Delta f_b/[f_{0,\min} + (N - 1)\,\Delta f_b]$$

That is, $k_{b,i}$ decreases with increasing N.

1.4. Noise Immunity of a Receiver

The radio signals reaching a receiving antenna are usually corrupted. This happens owing to the complex way in which radio waves propagate in the transmission medium and also because they contain an unwanted, or spurious, component (or components) superimposed on the wanted signal. Such spurious components are broadly called *radio noise* or, simply, *noise*. It may be produced by various causes, both natural and artificial. Since noise often hinders or even prevents normal signal reception, some of its forms are referred to as *interference*. Other forms are caused by disturbances in, say, the atmosphere and are called, by extension, *radio disturbances*.

The transmission medium may contain irregularities which cause signal energy absorption and scattering, multipath propagation, the Doppler frequency shift, and changes in wave polarization. This gives rise to signal fading, waveform mutilation, and intersymbol interference. These forms of noise are characterized by random variation in the complex transfer function of the medium and are called *multiplicative noise*. Ray delays due to multipath propagation and the Doppler shift are especially typical of radio links using long-distance tropospheric and ionospheric wave propagation.

A radio link may be treated as a nonstationary, linear system with random variables. Let the transmitted signal be

$$x(t) = \dot{X}(t)\exp(j\omega t)$$

where $\dot X(t)$ is the complex envelope. The received signal reflected by elementary scatterers in the scattering volume is

$$\dot z(t)=\sum_i \dot\mu_i \dot X(t-\Delta t_i)\exp[j(\omega+\Delta\omega_i)(t-\Delta t_i)]$$

where $\dot\mu_i = \mu_i(\Delta t_i, \Delta\omega_i)$ is the complex gain. Then the medium may be described by the transfer function $\dot H(\omega, t)$ found as the response to a signal of the form $\dot x(t) \equiv \exp(j\omega t)$:

$$\dot H(\omega, t)=\sum_i \dot\mu_i \exp[j(t\omega_i-\Delta t_i\omega-\Delta t_i\,\Delta\omega_i)]$$

The function $\dot{\mathrm{H}}(\omega, \mathrm{t})$ is a complex, nonstationary process. Therefore, in order to fully characterize multiplicative noise, one needs to know its multidimensional probability distribution.

Spurious waves which combine with the wanted signal form *additive noise.* It may be in-system and out-of-system, depending on whether its source is internal or external to a given communication system.

A portion of noise may be due to natural causes. Falling in this category are atmospheric noise, noise due to the thermal radiation of the Earth, and cosmic (or galactic) noise. Atmospheric noise is produced by lightning discharge in thunderstorms and by precipitation (rain, hail, snow or dust) hitting the antenna.*

Man-made noise or interference may be unintentional and intentional, commonly called jamming. Unintentional interference is usually produced by other radio transmitters whose signals fall within the desired or spurious reception channels, by the local oscillators of nearby receivers, and by industrial equipment, medical facilities, fluorescent lighting, car ignition systems, and other electrical and electronic equipment that radiates spurious signals.

Apart from all the above sources of noise, there remains internal noise originating in the electron devices and electric circuits of a receiver and associated devices. It is collectively known as *fluctuation noise.*

The ability of a receiver to stand up to the effects of noise is described as its *noise immunity.*

Any of the above forms of noise can find its way into the desired reception channel or occur outside it. Accordingly, there may be in-channel and out-of-channel noise or interference, respectively. Out-of-channel interference can be eliminated by frequency discrimination. In-channel noise mixed with the spectral components of the

* This kind of disturbance is often called precipitation static. See, for example, Reference Data for Radio Engineers, 5th edition. Howard W. Sams & Co, Inc., p. 27-4.— *Translator's note.*

wanted signal is far more difficult to combat. This is usually achieved by utilizing differences in the spectral, statistical and other properties of the signal and interference. Among other things, this purpose can be served by noise-immune forms of modulation, error-correcting codes, and special methods of signal processing at the receiver.

Quantitatively, noise immunity can be evaluated in terms of probability, power, and intelligibility criteria.

The probabilistic criterion is convenient in the case of discrete signals; it gives the mean probability of an elementary signal being mutilated

$$p = \sum_{i=1}^{m} p_i p(x_i)$$

where $p\ (x_i)$ = apriori probability of receiving the ith signal
p_i = probability of the ith signal being mutilated
m = size of the signal alphabet

If $p_i = p_0 = \text{const}$, which is the case with symmetrical binary channels using frequency-shift or phase-shift keying, then $p = p_0$. The quantity p_0 is a function of the signal-to-noise (or signal-to-interference) ratio. This ratio can be defined as

$$h^2 = E_s/\nu_{\text{int}}^2$$

where E_s is the signal energy and ν_{int}^2 is the interference spectral power density, or, in the case of a Gaussian noise as

$$h^2 = P_{\text{s}}/P_{\text{n}}$$

where P_{s} = mean power of the signal
P_{n} = mean power of fluctuation noise

The function $p_0\ (h^2)$ depends on the form of signal modulation, manner of reception, and the properties of the propagation medium. Its graphical representation is referred to as the *noise-immunity characteristic.*

The quantity p_0 does not give a complete idea about noise immunity. The probability p_{c} of a message being distorted and of a code combination being decoded with an error depends on the properties of interference and the manner of decoding. For telegraph channels, p_{c} may be 10^{-3} or 10^{-4}, whereas for data transmission systems it is usual to have p_{c} equal to or less than 10^{-6}. In element-by-element reception with a probability of false reception equal to p_0,

$$p_{\text{c}} = 1 - (1 - p_{0k})$$

where k is the number of symbols in the code combination.

The power criterion is more convenient when one has to deal with analog (continuous) signals. This may be the signal-to-noise power or rms voltage ratio at the receiver output for a specified signal-to-noise ratio at the input. However, it does not give a complete picture of signal and interference transmission through the receiving channel.

The articulation or intelligibility criterion is applicable to telephone channels. It is the percentage of speech units understood by a listener. The word 'articulation' is customarily used when the contextual relationships among the units of speech material are thought to play an unimportant role. In contrast, when the context is thought to play an important role in determining the listener's perception, this criterion is referred to as *intelligibility*.

1.5. Receiver Sensitivity

The sensitivity of a receiver is that characteristic which determines the minimum signal to which the receiver will respond. Quantitatively, it is defined at the minimum modulated-signal emf, V_{A0}, in a dummy receiving antenna, or the minimum field intensity, E_{A0}, or the minimum signal power, P_{A0}, at the receiver input. The minimum modulated-signal emf is used in the case of receivers operating in the LF through VHF bands. The minimum field intensity is considered when the receiver in question uses a magnetic or rod antenna. The minimum signal power is primarily used in assessing the sensitivity of UHF and SHF receivers.

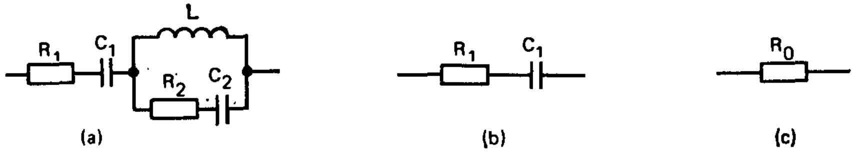
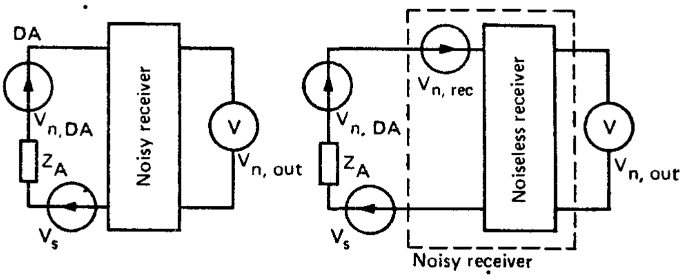


Fig. 1.3

Typical dummy (or artificial) antennas are two-terminal (one-port) networks whose averaged parameters are close to the likely parameters of a real antenna. An example of an open dummy antenna for LF through HF broadcast receivers is shown in Fig. 1.3*a*. Here, $R_1 = 50\ \Omega$, $R_2 = 320\ \Omega$, $C_1 = 125$ pF, $C_2 = 400$ pF, and $L = 20\ \mu$H. In the LF and MF bands, where X_L is low, use may be made of a simplified arrangement, such as shown in Fig. 1.3*b*. Finally, in the HF band, where X_L is high while X_{C1} and X_{C2} are low, the dummy antenna reduces to a single resistance, $R_0 = R_1 + R_2$ (Fig. 1.3*c*). The dummy antenna simulating a half-wave dipole is a 75-Ω resistor.

Sensitivity may be gain-limited, real, and threshold. The gain-limited sensitivity is typical of receivers with a relatively low gain, which receive strong signals so that interference can only negligibly affect the quality of reception. It is defined for a specified output power. For analog-signal (say, broadcast) receivers, one has to do with the maximum undistorted output power and normal output power. The former, $P_{s,m}$, is the maximum power which results from a 100%-modulated input signal, with nonlinear distortion not exceeding a certain level. The normal output power $P_{s,n}$ is that which results from a 30%-modulated input signal and amounts to 10% of $P_{s,m}$.

The real sensitivity of a receiver takes into account the effect of internal noise and is defined as the minimum input signal required to produce a specified output signal and signal-to-noise ratio, h^2_{out}. The threshold sensitivity of a receiver is the input signal level such that $h^2_{out} = 1$.



(a) (b)

Fig. 1.4

In summary, the sensitivity of a receiver is determined by its gain K, its internal noise level referred to the antenna input $V_{n, A\Sigma}$, and the required signal-to-noise ratio h^2_{out}. Let us take a closer look at how these factors affect the sensitivity of an AM receiver connected to an open dummy antenna. The receiver gain is

$$K = V_{s,out}/mV_{As} \tag{1.1}$$

where m = modulation factor

V_{As} = carrier-frequency rms voltage across the dummy antenna

Let V_{A0m} designate the value of V_{As} required to produce at the receiver output a voltage $V_{s,out}$. Then

$$V_{A0m} = V_{s,out}/mK \tag{1.2}$$

As is seen, the gain-limited sensitivity increases with increasing gain (V_{A0m} decreases).

In order to determine the real sensitivity, A_{A0r} (K), it is necessary to see how K affects the noise level at the receiver output, $V_{n,out}$. A real, noisy receiver connected to a noisy dummy antenna, DA (Fig. 1.4a) is replaced by a noiseless receiver containing a generator of receiver noise, $V_{n,\,rec}$, referred to its input, which along with the noise generator in the dummy antenna, $V_{n,\,DA}$, constitutes a generator of overall noise voltage, $V_{n,\,A\Sigma}$, referred to the dummy antenna (Fig. 1.4*b*), such that the rms voltage within the receiver bandwidth is

$$V_{n,\,A\Sigma} = (V^2_{n,\,DA} + V^2_{n,\,rec})^{1/2}$$

Since

$$V_{n,\,out} \approx KV_{n,\,A\Sigma}$$

it follows, in view of Eq. (1), that

$$V_{A0}/V_{n,\,A\Sigma} = V_{s,\,out}/mV_{n,\,out}$$

For a specified output signal-to-noise voltage ratio

$$h_{out} = (V_s/V_n)_{ont}$$

the signal-to-noise voltage ratio in the dummy antenna should be

$$h_A = V_{A0}/V_{n,\,A\Sigma}$$

Hence, the real sensitivity is

$$V_{A0r} \geqslant h_A V_{n,\,A\Sigma} \tag{1.3}$$

If we plot Eqs. (1.2) and (1.3) as shown in Fig. 1.5, the curves will intersect at point O representing the critical gain, K_{cr}. At $K < K_{cr}$, the gain is low, $V_{A0} > h_A V_{n,\,A\Sigma}$, $V_{s,\,out} > h_{out} V_{n,\,ont}$, and the real sensitivity V_{A0r} is gain-limited. At $K > K_{cr}$, $V_{A0} < h_A V_{n,\,A\Sigma}$ and, if $V_{A0} = V_{A0m}$, then

$$V_{s,\,out} < h_{out} V_{n,\,out}$$

which runs counter to the definition of real sensitivity. If one is to satisfy the equality

$$V_{s,\,out} = h_{out} V_{n,\,out}$$

one should raise V_{A0} to $h_A V_{n,\,A\Sigma}$. This means that real sensitivity is independent of K and is solely decided by receiver noise. As K rises above K_{cr}, the output signal voltage $V_{s,\,out}$ increases, as does $V_{n,\,out}$, whereas h_{out} remains unchanged.

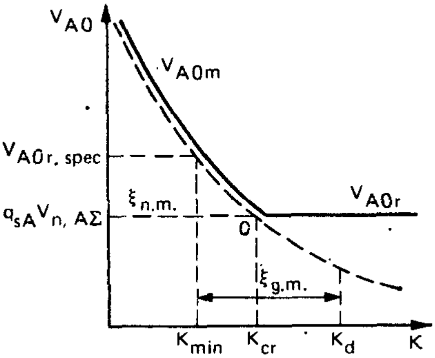


Fig. 1.5

If the real sensitivity is specified in advance to be equal to some value $V_{A0r,\,spec}$, it will be a good plan to design the receiver so that

$$V_{A0r,\,spec} = \xi_{n.m.} q_{sA} V_{n,\,A\Sigma}$$

where $\xi_{n.m.}$, approximately equal to 1 or 2, is the noise margin.

In service, the receiver gain might fall if V_{A0r} should turn out to be lower than has been specified. To avoid this, it is usual to provide some gain margin. Then the design gain is

$$K_d = \xi_{g.m.} K_{min}$$

where $\xi_{g.m.}$ greater than unity, is the gain margin.

As a way of protecting the final stage against overloading, the gain should be manually adjusted to be close to K_{cr}.

Receiver sensitivity is a function of frequency setting, and the nominal real sensitivity corresponds to the maximum value of V_{A0r}.

Depending on the quality class, the nominal real sensitivity for broadcast receivers ranges between 50 and 300 μV in the LF and MF bands, between 50 and 200 μV in the HF band, and between 3 and 30 μV in the VHF and UHF bands. For professional non-printing telegraph receivers operating in the HF band it may be as high as 0.1 μV, and for TV receivers the figure is anywhere between 200 and 500 μV.

1.6. Susceptibility of a Receiver to External Disturbance

The susceptibility of a receiver refers to its response to the disturbances acting through its antenna and other inputs. Receiver susceptibility is differently stated in terms of disturbance power, disturbance power flux density, or the electric or magnetic field strength associated with the disturbances. When assessing receiver susceptibility towards the noise and interference acting over supply, control and switching circuits, one takes into account interference voltage, current, and frequency.

If the receiver components have a high susceptibility to interference, this may affect the operating conditions and cause a reduction in the signal-to-noise ratio. Receiver susceptibility can be judged subjectively (such as when the operator passes judgement on the quality of telephone signal reproduction) · and objectively (such as when one determines the number of corrupted signal units in telegraphy).

While possessing susceptibility to interference, receiver components may themselves be sources of interference. Resistors are susceptible to interference that causes a rise in their resistance due to high-frequency heating and induces a stray emf (this is especially true of film and spiral wire-wound resistors). In turn, resistors produce thermal noise which can be a real nuisance in low-noise amplifiers. Inductors produce stray electromagnetic fields and, at the same time, are susceptible to interference. Capacitors are susceptible to nuclear, luminous and X-ray irradiation secondary to the ionization of their dielectric.

Current-carrying conductors set up electric and magnetic fields acting on other receiver components, and are at the same time susceptible to external fields. Semiconductor devices generate noise and, in turn, are susceptible to electromagnetic interference, gamma-rays, etc. Low-power transistors and low-capacitance diodes (with a junction capacitance of a fraction of a picofarad) can absorb electromagnetic energy, and this might cause them to fail.

Electromechanical switches usually have a low susceptibility to electromagnetic interference, except that which is strong enough to break down the contact gap. Faults in plug-and-socket connectors may be responsible for overheated contacts, sparking, arcing and, as a corollary, noise.

The receiver units may interact in several ways, each giving rise to detrimental effects of its own. Interaction via an electromagnetic field is the most common cause of interference which arises in, say, multi-turn inductors and fast-rise/fast-fall pulsers. Interaction by way of an electromagnetic field results in an interfering emf induced due to the capacitance existing between electric circuits. Radio-frequency radiation arises from hook-up wires and other structural components of a receiver to surrounding objects. Contact-potential interaction between metal structural components brings about changes in contact resistance, arc discharges, and contact deterioration as a result of galvanic and electrolytic corrosion. Luminous radiation arises when use is made of various photocells in which variations in the luminous flux cause changes in conductance or the generation of a voltage.

All of the above effects can adversely affect receiver performance, and appropriate measures need to be taken to mitigate them.

1.7. Noise in Receivers

Thermal noise. Any circuit possessing an ohmic resistance can be a source of thermal noise. The mean-square noise emf within the bandwidth of a receiver is given by the Nyquist equation

$$\overline{\mathrm{E}_{\mathrm{n}}^{2}} = 4kTBR \tag{1.4}$$

where k = Boltzmann's constant equal to 1.38×10^{-23} J deg^{-1}
T = absolute temperature of the circuit
kT = measure of thermal noise intensity in a resistance of 1 Ω per hertz of bandwidth
B = bandwidth in which the noise emf is measured
R = resistance of the circuit

Thermal noise is solely associated with resistance because it is produced by thermal agitation of electrons. Reactances (capacitive and inductive) are due to magnetic and electric fields where electron fluctuations are non-existent.

For purposes of analysis, it is customary to represent circuit noise by an equivalent noise voltage generator E_n, as shown in Fig. 1.6*a*, or by an equivalent noise current generator, as shown in Fig. 1.6*b*, where

$$I_n = (4kTBG)^{1/2} \qquad (1.5)$$

where G is the circuit conductance.

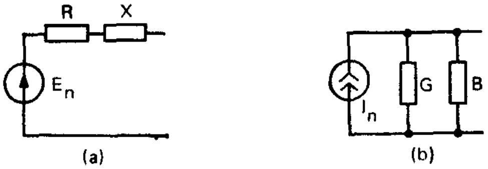


Fig. 1.6

In a tuned (or resonant) circuit, the noise source is the series loss resistance, r. The noise voltage across a parallel resonant circuit is Q times the noise emf. (The Q is the tuned-circuit quality factor or figure of merit; it is the reciprocal of the tuned-circuit damping factor, d). Then, in view of Eq. (1.5),

$$\bar{V}_{n,ckt} = QE_n = Q\,(4kTBr)^{1/2} = (4kTBR_0)^{1/2} \qquad (1.6)$$

At room temperature $T = T_0 = 293$ K,

$$V_{n,ckt} \approx (RB)^{1/2}/8$$

where $V_{n,ckt}$ is in microvolts, R is in kilohms, and B is in kilohertz. Here, $R_0 = Q^2r$ is the resonance resistance of the parallel resonant circuit.

In Eq. (1.5), applicable to cases where the noise source is represented by a current generator, $G_0 = 1/R_0$.

Antenna noise. A receiving antenna is the source of thermal noise associated with its loss resistance, and of the noise arising from the reception of radiation from outer space (galactic noise), the atmosphere, and the Earth. Thermal noise arising in an antenna plays a minor role, and the mean square noise emf due to external radiation can conveniently be evaluated by a relation similar to Eq. (1.4)

$$\overline{E}^2_{n,A} = 4kT_A\,Br_A \qquad (1.7)$$

where $T_A = t_A\,T$ is the noise temperature of the antenna, defined as the equivalent temperature at which thermal noise in r_A is the same as the actual antenna noise in its effect. It is convenient to regard T_A as a sum of components

$$T_A = T_g + T_{atm} + T_e$$

where T_g, T_{atm} and T_e are the noise temperatures associated with the galactic noise picked up by the antenna and with the effect of the atmosphere and the Earth. Also, T_A is a function of frequency, the radiation pattern of the antenna, and its orientation.

Noise due to amplifying devices. As will be recalled, the sources of internal noise in bipolar junction transistors (BJTs) are the thermal agitation of charge carriers in the base, emitter and collector;

fluctuations of emitter and collector current (shot noise); and random fluctuations as the emitter current divides between the electrodes (partition noise). The noise properties of BJTs are assessed in terms of noise parameters, namely noise resistance defined as

$$R_{\mathrm{n}} = (e/2kT)\,(h_{21\mathrm{e}}i_{\mathrm{e}}/|\,Y_{21}\,|^2) \approx 20h_{21\mathrm{e}}i_{\mathrm{e}}/\,|\,Y_{21}\,|^2 \qquad (1.8)$$

and the relative noise temperature of the input conductance defined as

$$t_{\mathrm{in}} = (1/G_{\mathrm{in}})\,[20i_{\mathrm{e}}\,(1 - h_{21\mathrm{e}}) + r_{\mathrm{b}}\omega^2C_{\mathrm{in}}^2] \qquad (1.9)$$

Here e = charge on an electron
$h_{21\mathrm{e}}$ = common-emitter current gain of the BJT
i_{e} = direct emitter current
r_{b} = base spreading resistance
ω = operating angular frequency
Y_{21}, G_{in}, C_{in} = BJT parameters

The noise properties of a BJT depend on supply conditions and frequency. Thermal resistance is tens of ohms, and the relative noise temperature of the input conductance is seldom greater than unity.

An equivalent noise circuit of a BJT, with its noise current or voltage generator placed at the input, may be depicted as shown in Fig. 1.7. The noise voltage generator

$$V_{\mathrm{n}} = (4kTBR_{\mathrm{n}})^{1/2} \qquad (1.10)$$

represents shot and collector-current partition noise. The noise current generator

$$I_{\mathrm{n,in}} = (4kT_{\mathrm{in}}G_{\mathrm{in}}B)^{1/2} \qquad (1.11)$$

accounts for thermal noise, shot noise, and partition noise in the base circuit. In Eq. (1.11), $T_{\mathrm{in}} = t_{\mathrm{in}}T$ is the noise temperature of the input conductance. The equivalent circuit in Fig. 1.7 applies to any of the three circuit configurations: common-emitter (CE), common-base (CB), and common-collector (CK) in which a BJT can be connected. The only difference is in the parameters defined by Eqs. (1.8) and (1.9).

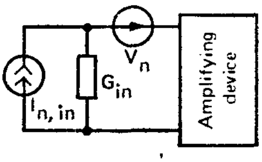


Fig. 1.7

The forms of noise associated with field-effect transistors (FETs) are thermal noise in the channel, shot noise in the gate, and thermal noise of the input conductance.

Thermal noise in the channel can be described in terms of noise resistance

$$R_{\mathrm{n}} = (0.6 \text{ to } 0.75)/S \qquad (1.12)$$

where S is the slope of the characteristic curve of the FET* called the transconductance.

Shot noise in the gate is markedly smaller than the thermal noise of the input conductance, and it is usually ignored. The relative noise

* Outside the USSR, the more common symbol is g_{m}.— *Translator's note.*

temperature of the input conductance G_{in} in a FET is

$$t_{in} \approx 1 \qquad (1.13)$$

An equivalent noise circuit of a FET is the same as the one shown in Fig. 1.7, where V_n and $I_{n,in}$ are defined by Eq. (1.10) and Eq. (1.11), respectively, subject to Eq. (1.12) and Eq. (1.13). The same equivalent circuit applies to a vacuum tube. For a triode tube, as an example, $R_n = (2 \text{ or } 3)/g_m$ and $1 < t_{in} < 5$.

The noise figure. The standard signal generator used to test a receiver is at the same time the source of noise defined by Eq. (1.4) or (1.5). Maximum power will be transferred from the generator to a completely matched load. This is known as the available generator output power

$$P_O = E_g^2/4R_g = I_g^2/4G_g \qquad (1.14)$$

According to Eqs. (1.4), (1.5) and (1.14), the available noise-output power is independent of the generator resistance

$$P_{O,\,g.n.} = kTB \qquad (1.15)$$

If the load is not perfectly matched, the noise-output power is given by

$$P_{g.n.} = \eta kTB \qquad (1.16)$$

Here, $\eta = P/P_O$ is the mismatch coefficient, and P is the power actually transferred to the load.

The noise properties of the signal generator are stated in terms of the ratio of the mean signal power to the mean noise power. In the four-terminal network (or the two-port) through which the signal passes from the generator the signal-to-noise power ratio is impaired because the two-port adds its own noise. The noise properties of the two-port are stated in terms of the *noise factor* which shows how many times the output signal-to-noise ratio decreases compared with the input signal-to-noise ratio

$$\begin{aligned} N &= (P_{s.g.}/P_{g.n.})/(P_{s,out}/P_{n,out}) \\ &= P_{n,out}/K_P P_{g.n.} \end{aligned} \qquad (1.17)$$

where

$$K_P = P_{s,out}/P_{g.n.}$$

is the *power gain*. The product $K_P P_{g.n.}$ is the noise-output power due to the signal generator. It is seen that the noise factor is defined as the ratio of the total noise output power to its fraction due to the signal generator noise. Therefore, Eq. (1.17) may be re-written as follows:

$$N = \frac{P_{g.n}K_P + P_{n,int}}{P_{g.n}K_P} = 1 + P_{n,\,int}/P_{g.n}K_P \qquad (1.18)$$

where $P_{n.\,int}$ is the internal noise output power of the two-port.

It should be clearly realized, however, that the noise factor evaluates only the linear part of the receiver, that is, up to the detector. The noise factor of a passive two-port (such as an antenna feeder) properly matched to both the signal generator and the load is a function of the power gain

$$N = 1/K_P \tag{1.19}$$

Owing to losses in passive circuits, $K_P < 1$ and $N > 1$.

For comparison of noise and the signal at the output, it is convenient to refer all of the noise to the input, assuming that the receiver proper is free from noise and only amplifies input noise. As follows from Eq. (1.18), two-port noise referred to input is

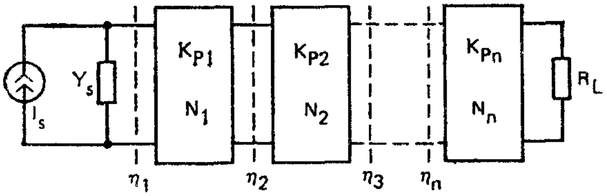


Fig. 1.8

$$P_{n,in} = P_{n,int}/K_P = (N - 1)\, P_{n,g} \tag{1.20}$$

or, if one proceeds from Eq. (1.16),

$$P_{n,in} = (N - 1)\, \eta kTB \tag{1.21}$$

Let us find the noise factor for a linear network composed of a series connection of several two-ports, such as amplifiers (Fig. 1.8). Each two-port can be described in terms of its power gain K_{Pi} and its noise factor N_i. Suppose that the mismatch coefficients η_1, η_2, . . ., η_n at the junctions of the two-ports are known. According to Eq. (1.17),

$$N = 1 + \left(\sum_{i=1}^{n} P_{n,\,out,\,i}\right) \Big/ P_{n,\,g,\,out} \tag{1.22}$$

where

$$P_{n,g,out} = P_{g.n.}K_{P1}K_{P2} . . . K_{Pn}$$

is the generator noise output power. In view of Eq. (1.16),

$$P_{n,g,out} = \eta_1 kTBK_{P1}K_{P2} . . . K_{Pn} \tag{1.23}$$

In agreement with Eq. (1.21), the noise output power of the first two-port is

$$P_{n,out,1} = (N_1 - 1)\, \eta_1 kTBK_{P1}K_{P2}K_{P3} . . . \tag{1.24}$$

The noise of every next two-port is amplified by all stages except the preceding ones. Therefore, similarly to Eq. (1.23).

$$\left.\begin{aligned} P_{\mathrm{n,\,out}\,2} &= (N_2 - 1)\,\eta_2 kTBK_{P2}K_{P3}K_{P4}\ \ldots \\ P_{\mathrm{n,\,out}\,3} &= (N_3 - 1)\,\eta_3 kTBK_{P3}K_{P4}K_{P5}\ \ldots \\ \ldots&\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned}\right\} \qquad (1.25)$$

On substituting Eqs. (1.23) through (1.25) into Eq. (1.22), we get

$$N = N_1 + \frac{\eta_2}{\eta_1}\frac{N_2 - 1}{K_{P1}} + \frac{\eta_3}{\eta_1}\frac{N_3 - 1}{K_{P1}K_{P2}} + \ldots \qquad (1.26)$$

Apart from the noise factor, use is widely made of the *noise temperature* defined as

$$T_{\mathrm{n}} = (N - 1)\,T \qquad (1.27)$$

It characterizes the own noise of a two-port referred to its input. This quantity is a thermal equivalent of two-port noise and shows how many degrees the dummy antenna should be heated in order that its noise output power could equal the own noise of the two-port. It is convenient to apply the concept of noise temperature to low-noise amplifiers which have a noise factor of close to unity. For example, assuming that $N = 1.1$, it follows from Eq. (1.27) that $T_{\mathrm{n}} \approx$ 30 K. According to Eqs. (1.26) and (1.27), the noise temperature of a multistage device is

$$T_{\mathrm{n}} = T'_{\mathrm{n}1} + \frac{\eta_2}{\eta_1}\frac{T_{\mathrm{n}2}}{K_{P1}} + \frac{\eta_3}{\eta_1}\frac{T_{\mathrm{n}3}}{K_{P1}K_{P2}} + \ldots$$

The noise factor and noise temperature are mainly decided by the properties of the front-end two-ports. The effect of the succeeding stages decreases in proportion to the power gain of the preceding ones. For the noise factor to be low, tbe early stages of a receiver should be made to have a low noise level and a high power gain.

Noise sensitivity of a receiver. This quantity is defined as the minimum signal power or emf in the antenna required to ensure a specified signal-to-noise ratio at the output of the linear portion of the receiver:

$$h_{\mathrm{out}} = P_{\mathrm{s,out}}/P_{\mathrm{n,out}} \qquad (1.28)$$

This takes into account antenna noise as given by Eq. (1.7), feeder noise, and receiver noise. At perfect match, the antenna noise power transferred to the feeder input will, in agreement with Eqs. (1.7) and (1.14), be defined as

$$P_{\mathrm{n,A}} =.\, kT_{\mathrm{A}}B$$

Let the feeder gain be denoted as K_{PF}. Then the noise power at the receiver input will be given by $P_{\mathrm{nA}}K_{\mathrm{PF}}$. Assuming that the feeder is perfectly matched, the feeder noise power at the receiver input will, in agreement with Eqs. (1.19), (1.20) and (1.15), be given by

$$P_{\mathrm{nF}} = kTB\,(N_F - 1)K_{\mathrm{PF}} = kTB\,(1 - K_{\mathrm{PF}}) \qquad (1.29)$$

where $N_F = 1/K_{\mathrm{PF}}$ is the noise factor of the feeder. If N is the noise factor of the receiver, then the receiver noise power referred to its input is given by Eq. (1.21), and the overall noise power at the output of the linear section with a power gain equal to K_{P} and $\eta = 1$ will be

$$\begin{aligned} P_{\mathrm{n,out}} &= (P_{\mathrm{n,A}}K_{\mathrm{PF}} + P_{\mathrm{nF}} + P_{\mathrm{n,xvr}})\,K_{\mathrm{P}} \\ &= kTB\,[K_{\mathrm{PF}}\,(t_{\mathrm{A}} - 1) + N]\,K_{\mathrm{P}} \qquad (1.30) \end{aligned}$$

where $t_{\mathrm{A}} = T_{\mathrm{A}}/T$ is the relative noise temperature of the antenna.

As follows from Eq. (1.28), if one is to obtain the desired quality of reception, the signal power at the output of tho linear section should be

$$P_{\mathrm{s,out}} = q_{\mathrm{out}}P_{\mathrm{n,out}}$$

Therefore, the signal power in the antenna, defining the receiver sensitivity, is given, in agreement with Eq. (1.30), by

$$P_{\mathrm{s,A}} = P_{\mathrm{s,out}}/K_{\mathrm{PF}} = kTBq_{\mathrm{out}}\,(N/K_{\mathrm{PF}} + t_{\mathrm{A}} - 1)$$

Accordingly, the signal emf in the antenna, defining noise sensitivity is

$$E_{\mathrm{A}} = [4kTr_{\mathrm{A}}Bq_{\mathrm{out}}\,(N/K_{\mathrm{PF}} + t_{\mathrm{A}} - 1)]^{1/2}$$

1.8. Frequency Selectivity of a Receiver

In contrast to noise immunity and sensitivity, selectivity cannot be stated in terms of a single number. The point is that because the electron devices used in a receiver are nonlinear, the effects arising as a signal and interference pass through together might differ from those that would take place if the signal and interference were received separately. The most troublesome effects are blocking interference, cross-talk interference, and intermodulation interference.

Blocking interference manifests itself as a reduction in the signal amplitude at the amplifier output due to the action of a strong interference; as a result, interference stands out more prominently. Cross-talk interference has as one of its consequences the fact that the modulation of an undesired carrier is transferred onto the desired carrier. Intermodulation interference refers to the production of interfering frequencies which are close to the signal frequency and correspond to the sums and differences of the fundamentals and harmonics of two or more strong interfering signals differing in frequency and transmitted through the device simultaneously.

It is most of all difficult to mitigate the interference at frequencies adjacent to the desired reception channel. Therefore, adjacent-channel selectivity is of primary importance.

In the simple single-signal method, selectivity is described by the frequency response of the receiver to a single low-level harmonic signal which is applied to the receiver input and causes no nonlinear effects. The gain of the r.f. section, K, is a function of frequency, and is a maximum, K_0, at f_0, which is the resonant (or resonance) frequency. The ratio $\sigma = K_0/K$ shows how much unwanted signals or interference is attenuated. The function σ (f) is plotted in the form of what is known as the *selectivity*, *resonance* or *frequency response curve*, as shown in Fig. 1.9. Frequency f or the amount off resonance, or detuning, $\Delta f = f - f_0$, is plotted as abscissas, and σ is plotted as ordinates in either relative units (ratios) or in decibels. In the latter case,

$$\sigma \text{ (dB)} = 20 \log_{10} (K_0/K)$$

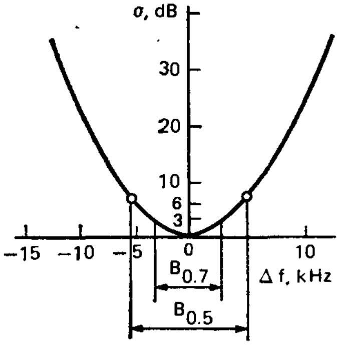


Fig. 1.9

Instead of σ, one may use the relative gain defined as $K/K_0 = 1/\sigma$.

The ideal from the viewpoint of selectivity is a rectangular response curve in which case σ (dB) is equal to zero within the bandwidth and tends to infinity outside the bandwidth. In real receivers, σ is non-zero and never constant within the bandwidth. Because of destabilizing factors, a need may sometimes arise to spread the bandwidth in excess of the width of the signal spectrum. This impairs selectivity, however, and it is important to ensure frequency stability of the receiver. When, on the other hand, strong interference is likely to fall within the bandwidth along with the desired signal, it is warranted to reduce the bandwidth and to modify the shape of the frequency response. Then the impairment in the quality of signal reproduction due to a limited signal spectrum will be less significant than the improvement in the quality of reproduction due to the reduced effect of undesired signals. For this reason, receivers have sometimes provisions for bandwidth control, that is, variable selectivity.

Ordinarily, the bandwidth B is measured between 3-dB points, that is, at $\sigma = 3$ dB, or at $K/K_0 = 0.7$. The bandwidth at 0.7 level, $B_{0.7}$, corresponds to a frequency band within which the gain is $K \geqslant 0.7\ K_0$ or 3 dB down, and the midband frequency is the tuning, or resonance, frequency f_0. Sometimes the bandwidth is measured between 6-dB points, that is at $\sigma = 6$ dB, and is termed the half-power bandwidth, $\mathrm{B}_{0,s}$.

How close the real response curve comes to the ideal one is stated in terms of the *fractional bandwidth** μ_γ, where γ is the level at which

* It is often expressed as a percentage rather than a fraction, and is then termed the percentage bandwidth.— *Translator's note.*

the measurement is done, or in terms of the mean rolloff or slope of the response curve S_{fr}, or in terms of the relative attenuation or selectance σ_Δ at a specified amount Δf_0 off the desired frequency f_0.

Thus, the fractional f bandwidth is $\mu_\gamma = B_{0.7}/B_\gamma$ where $B_\gamma > B_{0.7}$ is the receiver bandwidth at $\gamma < 0.7$. Its reciprocal, $\xi_\gamma = 1/\mu_\gamma$, shows how much the bandwidth at the γ level exceeds the one at the 0.7 level. Selectivity improves progressively more as μ_γ approaches unity.

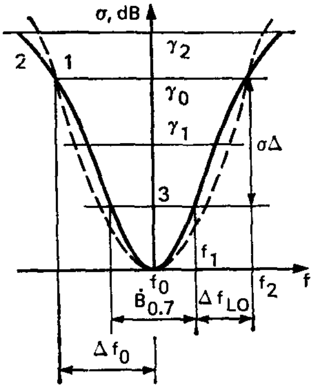


Fig. 1.10

Different receivers may have the same μ_γ at the specified value of γ, but it is not enough to assert that they have the same selectivity. Referring to Fig. 1.10, receivers *1* and *2* have the same $\mu_{\gamma 0}$, but at $\gamma_1 < \gamma_0$, $\mu_{\gamma_1}^{(1)} < \mu_{\gamma 2}^{(2)}$, which means that the frequency response of receiver *1* is closer to the square one than that of receiver *2*. At $\gamma_2 > \gamma_0$, the reverse is true. Therefore, it is usual to invoke the rolloff, or the slope, of the frequency response

$$S_{\text{fr}} = \left| \frac{\sigma_2 - \sigma_1}{f_2 - f_1} \right| \qquad (1.31)$$

measured in decibels at 1 kHz. Ordinarily, one chooses $\sigma_1 = 3$ dB, that is, $S_{\text{fr}} = (\sigma - 3)/\Delta f_1$, where Δf_1 is the amount of detuning relative to the edges of the bandwidth. It follows from Eq. (1.31) that, given the same frequency response, S_{fr} will have different values at different amounts off resonance.

The relative attenuation or selectance σ_Δ is given at some specified amount of detuning, Δf_0, on either side of the desired frequency f_0. For example, in testing LF and HF broadcast receivers, it is often assumed that $\Delta f_0 = 9$ kHz, which corresponds to the width of the allocated frequency band. Then frequencies $f_0 \pm 9$ kHz will be adjacent-channel carrier frequencies, and the respective values of σ_Δ will give what is called the *selectance against the next channel* or the *adjacent-channel attenuation* (ACA).

Apart from the selectance against the next channel, superhet receivers are characterized in terms of selectance against spurious responses or channels, stated as image rejection, σ_{im}, and intermediate-frequency rejection, σ_{if} (see Sec. 1.2). The values of σ_{im} and σ_{if} are decided by the selectivity of the preselector.

For a more accurate assessment of the effect produced by unwanted signals, one determines multi-signal selectivity. It takes into account the nonlinear effects arising due to the simultaneous action of a desired signal and strong interference. There may be several kinds of multisignal selectivity, but those used most often are two-signal and three-signal selectivity.

In assessing selectivity by the two-signal method, one uses two test generators, G_s and G_{int}, which feed to the receiver input a modulated signal voltage $V_{A,s}$ and a modulated interference voltage $V_{A,int}$ through a standard dummy antenna. These are rms r.f. voltages which, for AM receivers, are modulated 30% ($m_s = m_{int} = 0.3$) at 400 Hz ($F_{m,s}$) and 1000 Hz ($F_{m,int}$), respectively. To begin with, the signal frequency is set at $f_s = f_0 =$ const, G_{int} is turned off, and, while holding $V_{A,s}$ at a constant value, one notes the voltage at the receiver output, $V_{s,out}$, also held at a constant value. Then G_{int} is turned on, f_{int} is varied, and $V_{A,int}$ is adjusted so as to obtain an output signal-to-noise voltage ratio

$$h_{out} = (V_s/V_{int})_{out}$$

providing for the specified quality of signal reception. The value of $V_{A,int}$ thus obtained is then taken as the allowable one, $V_{A,int}$. The ratio $V_{A,int,allowable}/V_{A,s}$ as a function of f_{int} or of detuning $\Delta f_{int} = | f_{int} - f_0 |$ at a specified h_{out} gives a measure of receiver selectivity.

In assessing selectivity by the three-signal method, one uses three test generators G_s, G_{int1} and G_{int2} which feed to the receiver input voltages simulating the desired signal at frequency f_s and two interfering signals at frequencies f_{int1} and f_{int2}. Frequency f_{int1} is taken to lie above or below $f_s = f_0$ and corresponds to the adjacent-channel frequency, whereas $f_{int2} = 2f_{int1} - f_s$. Owing to the nonlinearity of the r.f. section, what are known as intermodulation frequencies are produced, which include one equal to $2f_s - f_{int2}$ and lying very close to f_s, for which reason this interference cannot be filtered out in the succeeding stages. In more detail, intermodulation will be discussed in Chap. 8.

1.9. The Dynamic Range of a Receiver

In actual service conditions, the desired and interfering signals may vary in amplitude by as much as 80 dB or even more. An excessive increase in signal amplitude might overload the r.f. section to a point where the output signal would be severely distorted. The ratio of the maximum allowable input signal voltage, $V_{A,allowable}$, to the receiver sensitivity, V_{A0}, defines the dynamic range of the receiver

$$D = 20 \log_{10} (V_{A,allowable}/V_{A0})$$

For broadcast receivers, D extends from 40 to 60 dB; for point-to-point receivers the figure is 60 to 80 dB or even more. The dynamic range can be extended through the use of electron devices whose volt-ampere characteristic has an extended linear region, and automatic gain control.

1.10. Fidelity

Fidelity refers to the degree to which the receiver accurately reproduces at its output terminals the modulation possessed by the received wave. Quantitatively, fidelity is assessed in terms of distortion, that is, changes in the shape of the output signal relative to the modulating function. In turn, distortion can be assessed by a spectral method and by direct comparison. The spectral method is based on comparing the spectra of the output wave and the modulating voltage of the input signal. Distortion can be linear and nonlinear.

Linear distortion is due to the inertia of the components used in the receiver and is not accompanied by the production of previously nonexistent components in the output signal spectrum; it is independent of both the input signal and the depth of modulation. There may be *amplitude distortion* and *phase distortion*. The former manifests itself as changes in the relative amplitudes of the spectral components; the latter arises from differences in time delay between these components.

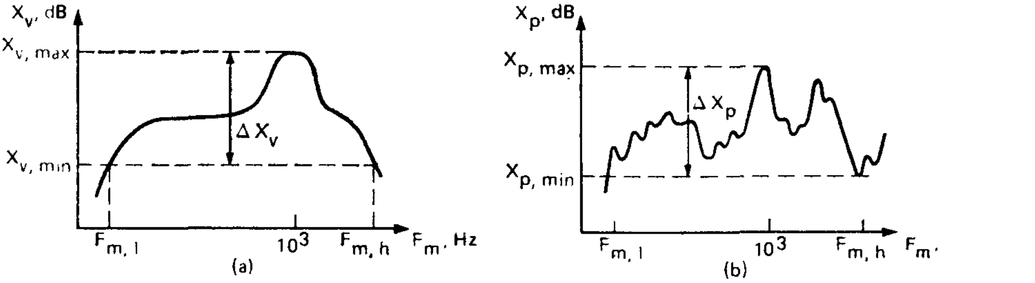


Fig. 1.11

Amplitude distortion is evaluated in terms of voltage or sound-pressure fidelity. The voltage fidelity X_v relates the receiver output voltage, $V_{s,out}$, to the modulation frequency F_m of the input signal, with the tuning frequency, input-signal amplitude and depth of modulation held constant. As a rule, $V_{s,out}$ is normalized in terms of the output signal voltage $V_{s,out,mod}$ with F_m equal to 400 or 1000 Hz

$$X_v = 20 \log_{10} (V_{s,out}/V_{s,out,mod})$$

A typical voltage fidelity curve is shown in Fig. 1.11*a*, where $F_{m,l}$ and $F_{m,h}$ are the lower and upper modulation frequencies. An ideal fidelity curve is a straight line at the 0-dB level. A real fidelity characteristic shows a cut at the lower and upper audio frequencies and a boost at the midband frequencies. The low-frequency cutting is due to interstage coupling capacitors or transformers. The upper-frequency cutting is due to the modulation-frequency (most often audio-frequency) and radio-frequency sections. In the modulation frequency section, the main 'culprit' is the shunting effect of the input and output capacitances of the electron devices and the wiring capacitance. In the r.f. section, the key role is played by the limited bandwidth.

Variations in voltage fidelity are stated in terms of a quantity which is defined as

$$\Delta X_{\mathrm{v}} = X_{\mathrm{v,max}} - X_{\mathrm{v,min}}$$

and is expressed in decibels. Quite often, it is admissible for ΔX_{v} to be less than or equal to 6 dB.

The voltage fidelity characteristic does not take into account the amplitude distortion contributed by the end device. This is the reason why the quality of broadcast receivers is stated in terms of the sound-pressure fidelity characteristic. It relates the sound pressure developed by the speaker, P_{sound}, to the modulation frequency F_{m} of the input signal, on the provision that the frequency to which the receiver is tuned is held constant, there is no overloading in any of its stages, and the distance to the point of measurement is maintained constant (equal to 1 m). Also, P_{sound} is usually normalized in terms of the sound pressure, $P_{\mathrm{sound,mod}}$ at the modulation frequency $F_{\mathrm{m}} = 400$ or 1000 Hz:

$$X_{\mathrm{p}} = 10 \log_{10} (P_{\mathrm{sound}}/P_{\mathrm{sound,mod}})$$

A typical sound-pressure fidelity curve is shown in Fig. 1.11*b* . The many peaks seen in the plot are due to the mechanical resonances of the speaker diffusor. Variations in the curve between $F_{\mathrm{m,l}}$ and $F_{\mathrm{m,h}}$ are stated in terms of the quantity

$$\Delta X_{\mathrm{p}} = X_{\mathrm{p,max}} - X_{\mathrm{p,min}}$$

For broadcast receivers, $\Delta X_{\mathrm{p}} \leqslant 14$ to 18 dB.

Phase distortion is evaluated in terms of the envelope delay characteristic: To begin with, phase shifts $\Delta\varphi$ at frequencies f within the bandwidth of the receiver are measured at a constant modulation frequency, $F_{\mathrm{m}} = \mathrm{const}$, following which one finds the envelope delay time

$$\tau_{\mathrm{d}} = \tau - \Delta\varphi/2\pi F_{\mathrm{m}}$$

A measure of phase distortion is given by

$$\Delta X_{\mathrm{ph}} = \tau_{\mathrm{d,max}} - \tau_{\mathrm{d,min}}$$

Nonlinear distortion gives rise to previously nonexistent components in the spectrum of the receiver output wave; they are functions of the input signal level and the depth of modulation. It is stated in terms of the *harmonic factor*

$$k_{\mathrm{h}} = (V_2^2 + V_3^2 + \, . \, . \;)^{1/2}/V_1$$

or the *nonlinear distortion factor*

$$k_{\mathrm{nl}} = [(\ldots V_2^2 + V_3^2 + \ldots)/(V_1^2 + V_2^2 + V_3^3 + \ldots)]^{1/2}$$

where V_1 are the rms harmonic voltages of the modulation frequency. Obviously, $k_{\mathrm{nl}} = k_{\mathrm{h}}/(1 - k_{\mathrm{h}}^2)^{1/2}$, and at $k_{\mathrm{h}} \leqslant 0.2$ it may be taken that $k_{\mathrm{nl}} \approx k_{\mathrm{h}}$*.

In aural reception, only amplitude and nonlinear distortion is essential; small phase distortion is passed unnoticed. In visual reception, as in television, both amplitude and phase distortion are important. In the latter case, the distortion is evaluated by watching the signal waveform on an oscilloscope, and the modulating function is usually a step change in voltage. The output voltage waveform is shown in Fig. 1.12 where $V_{\mathrm{s,out}}$ is normalized in terms of the steady-state signal voltage, $V_{\mathrm{s,ss}}$. The interval of time required for the leading edge of the waveform to rise from 0.1 to 0.9 of its peak amplitude is called the *rise time*, τ_{r}. The amount of time required for the leading edge to rise to half the steady-state signal value is known as the *delay time*, τ_{d}. The intensity of the damped oscillations, commonly called *ringing*, in the output waveform is evaluated in terms of Δ_i**. The delay time, τ_{d}, is mainly determined by the rate of roll-off, or slope, of the frequency response, S_{fr}: the delay time increases with increasing S_{fr}. The delay as such is not a sign of distortion in a message, but instability in the delay time might lead to untoward consequences, as in equipments which utilize the difference in the time of arrival of signals over different paths.

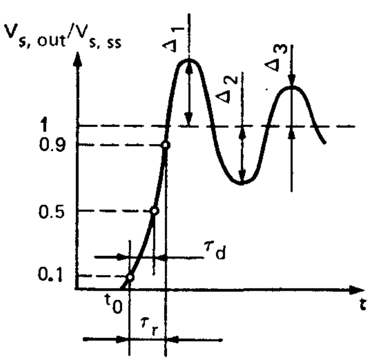


Fig. 1.12

The rise time, τ_{r}, is an approximate reciprocal of the receiver bandwidth, $\tau_{\mathrm{r}} \approx 1/B_{0.7}$. In TV receivers, an increase in the rise time causes a blur in the outline of the image.

* In fact, both are called either the *distortion factor* or the *total harmonic distortion* outside the USSR. See "Reference Data for Radio Engineers", Howard W. Sams & Co., p. 17-12, or "Hewlett-Packard Electronic Instruments and Systems", 1976, p. 436 — *Translator's note.*

** The initial transient response, labelled Δ_1 in our case is commonly referred to as the overshoot. More accurately, it is the amplitude of the first excursion of a wave beyond the 100% amplitude level.— *Translator's note.*

The magnitude of ringing, Δ_1, depends on the shape of the frequency response. With an ideal, 'square' frequency response, the theoretically computed ringing is such that $\Delta_1 = 9\%$, $\Delta_2 = 5\%$, and $\Delta_3 = 3\%$. In TV receivers, the first excursion, or *overshoot*, will do no harm if it is not over 5%; if it is greater than that, the quality of image reproduction will be seriously impaired. The second excursion would cause the image to double; to avoid this, it is important that it should not be greater than 2%. In radar receivers, heavy ringing and, especially, large overshoots may cause false targets to appear on the scope.

1.11. Receiver Controls

Apart from tuning controls, receivers are provided with manual and automatic controls and adjustments, namely:

— an automatic gain control to prevent overloading and to provide normal operating conditions for the end device;

— an automatic bandwidth control to mitigate adjacent-channel interference;

— an automatic local-oscillator control;

— antenna orientation and radiation-pattern switching controls;

— an antenna polarization switching control.

Automatic controls provide for no-search (pushbutton) tuning by setting up the frequency code or a fixed frequency number on the control panel; return to a previous frequency setting after operation on a new setting; retention of the frequency setting and operating conditions existing at the time of turn-off so that they can be restored when the receiver is turned on again; coded entry of preset frequencies into memory and their read-out after prolonged storage; spatial position control of a ferrite-rod antenna; selection of reception modes (mono-stereo-quadraphony, etc.); wireless remote control of a receiver, turntable, tape recorder, TV set, and the like, with all relevant operational data presented on the display of an electronic control panel.

Receivers used at radio centres have each a great number of controls and adjustments, especially in cases where use is made of the frequency-allocation service and of the data supplied by real-time analysis of the interference situation for adaptive control of receiver and system parameters. In such cases, receivers incorporate microcomputers.

1.12. Receiver Design and Performance Characteristics

These characteristics include performance stability, ergonomicity, reliability, maintainability, power requirements, weight and size, cost, mobility, and some others.

Stability refers to the ability of a receiver to preserve, while in service, its basic performance characteristics within specified limits in the face of variations in supply voltages, temperature, air humidity and pressure, exposure to mechanical, radioactive and chemical factors. Of special importance is the accuracy with which the frequency can be set and maintained and the gain stability.

The accuracy of frequency setting is stated in terms of the error, Δf_t, with which a receiver can be tuned to the desired frequency in the absence of a signal. The desired frequency is set and read by means of electronic digital read-out devices or tuning dials. The latter may be open (so that all of the scale is visible to the operator) or concealed (so that only a part of the entire scale is accessible for observation), the vernier type without magnification or the optical type with lense for projection magnification, etc. Furthermore, scales may be continuous or discrete.

Given continuous tuning and a continuous scale, an important considerations is the linear dimension, l, of the scale. Furthermore, tuning dials are characterized in terms of the frequency density of their scales, k_d, which gives the number of kilohertz per millimetre of scale length:

$$k_d = \Delta F_b k_{nl} / \eta l$$

where ΔF_b = width of the frequency band within which the receiver is tuned with the aid of the tuning dial

η = magnification of the optical system ($\eta \geqslant 1$)

k_{nl} = scale nonlinearity factor (ordinarily, $k_{nl} \approx 1.03$ to 1.1)

The value of k_d is related to Δf_t as $k_d = k_{va}\,\Delta f_t$, where k_{va} is the operator's visual acuity found under normal conditions of scale observation from a distance of 25 cm and at an angle of vision of 1 minute of arc (as a rule, $k_{va} \approx 5$ to 10). Hence,

$$\Delta F_b = k_{va} \eta l \Delta f_t / k_{nl}$$

In consequence, the width of a frequency band should be decreased and the number of bands should be increased with decreasing linear dimension of the scale and decreasing accuracy of frequency setting.

In receivers operating on preset frequencies, the desired frequencies are set with the aid of digital read-out devices, and the value of Δf_t is decided by the stability of the reference crystal oscillator. The entire frequency range is divided into bands so as to obtain the desired maximum-to-minimum frequency ratio within each band with the aid of a chosen tuning control. In point-to-point HF receivers, the preset frequencies are spaced 10 or 100 Hz apart and it usually takes from 0.2 to 1.2 seconds to move from one frequency to another.

It is often assumed that the limit of error in frequency setting should not exceed $\Delta f_0 = 0.1$ to 0.3 of the bandwidth. It is decided by the fractional stability of the local oscillator, defined as

$$\delta f_{\mathrm{LO}} \approx \Delta f_0/f_{\mathrm{LO}}$$

Since, however, f_{LO} is approximately equal to f_0, it follows that

$$\delta f_{\mathrm{LO}} \approx \delta f_0 = \Delta f_0/f_0$$

For MF and LF communications and broadcast receivers the required stability ($\delta f_0 \approx 10^{-3}$, or one part in a thousand) is provided by parametric stabilization. At the higher frequencies, resort is made to thermostatted or unthermostatted crystal stabilization ($\delta f_0 \approx \approx 10^{-4}$ to 10^{-6}, or from one part in ten thousand to one part in a million). In point-to-point receivers incorporating frequency synthesizers, frequency stability is one part in 10^7 to 10^8.

Requirements for gain stability in broadcast receivers are rather lenient. In design, the overall gain K is chosen with a margin (so that it is 1.5 to 2 times the value obtained by calculation). The excess gain is eliminated by automatic and manual controls. In special-purpose receivers (such as field-strength meters, interference analyzers, and the like) the need exists for high gain stability which is ensured through the use of feedback, periodic calibration against test signals, regulated power supplies etc.

Receiver ergonomicity refers to a set of properties essential for proper dynamic man-receiver interaction under specified service conditions. It is assessed in terms of such factors as visual perception of the panels by the operator (illumination and colour of the dials, light contrast, field of vision, etc.), aural perception (intensity and quality of sound, masking of sound signals by the background noise in the room), tactile perception (the effort applied by the operator in manipulating the controls), coordination of control motions (in local and remote control, number and type of controls, rate and accuracy of control setting and the like), receiver service conditions (range of air temperature and humidity, comfort of the operator's station).

Reliability is related to the mechanical and electrical strength of the receiver. For point-to-point receivers, the mean time to failure is thousands of hours in the temperature range from —10° to +50°C.

Maintainability has to do with ease of access to the various units and components for inspection, repair or replacement.

Power requirements depend on the type of electron devices and the end device used. Broadcast receivers can draw their power from either local supplies or an a.c. supply line. Point-to-point receivers operate off an a.c. supply line.

The weight and size of a receiver depend on its intended use and quality. Depending on the kind of components used, point-to-point receivers of medium quality may weight anywhere between 40 and 60 kg and take up a volume of 15 to 90 dm³, whereas high-quality point-to-point receivers tip the scales at 15 to 370 kg and measure 30 to 720 dm³ in volume.

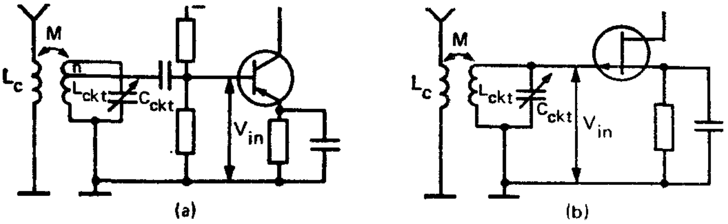
It is never possible to satisfy all the requirements to the same degree because some of them are conflicting. For example, the use of several tuned circuits in the input stages entails an impairment in receiver sensitivity. Therefore, when designing a receiver, one seeks a compromise which would ensure local optimization.

Chapter Two

Input Circuits of Receivers

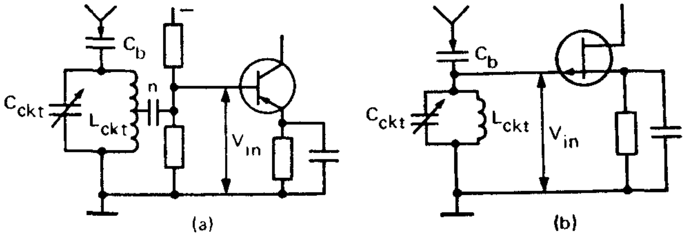
2.1. Purpose and Characteristics

The purpose of the input circuits is to pass the received signal from the antenna to the succeeding circuits and to effect preliminary interference suppression. The input circuit is usually a four-terminal (or two-port) network containing one or several frequency-selective sections (notably, resonant or tuned circuits) which extract the desired signal from the incoming waves. Most often, the input circuit has one tuned circuit. Two or more tuned circuits are only used when special requirements are to be met for selectivity.



(a) (b)

Fig. 2.1



(a) (b)

Fig. 2.2

Figure 2.1 shows an input circuit in which the antenna is coupled to the tuned circuit $L_{ck}{:}C_{ckt}$ by a transformer (a *transformer-coupled input circuit*). Figure 2.2 shows a *capacitive-coupled input circuit*, and Fig. 2.3, an input circuit with *tapped-coil coupling*.

The next stage can be tapped on all or a part of the tuned-circuit coil in the input circuit, depending on the input impedance of that stage. A BJT will usually be tapped down on the tuned-circuit coil

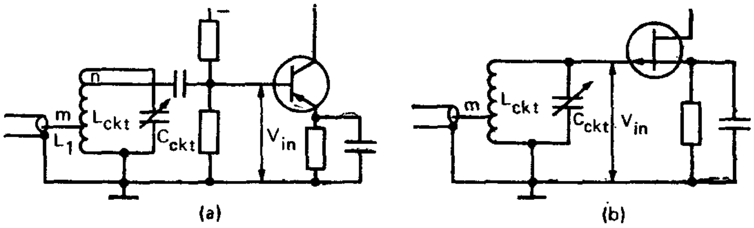


Fig. 2.3

because it has a low input resistance. A FET may be tapped to all of the tuned-circuit coil.

Figure 2.4 shows a double-tuned input circuit, that is, one using two tuned circuits. The first tuned circuit is transformer-coupled to the antenna. The tuned circuits are coupled by capacitor C_{c1} (internal capacitive coupling) and capacitor C_{c2} (external capacitive coupling). A double-tuned input circuit offers a means to obtain a nearly square frequency response, that is, to enhance selectivity.

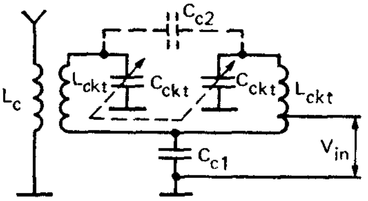


Fig. 2.4

The principal variables that characterize the input circuit of a receiver are as follows:

— gain, that is, the ratio of the signal voltage at the input to the next stage, V_{in}, to the emf (open-circuit voltage) in the antenna, E_A; in the case of a ferrite-rod antenna, it is the ratio to the signal field intensity;

— bandwidth, that is, the total width of the frequency band over which variations in the gain remain within specified limits;

—selectivity, that is, the degree to which the gain at a given amount off resonance (or away from the midband frequency), K, is reduced in comparison with its value at resonance (or at the midband frequency), K_0, that is, $\sigma = K_0/K$. The input circuit and the r.f. amplifier provide the desired discrimination, or selectance, between the desired signal and signals of other carrier frequencies and preliminary filtration of interference;

— frequency coverage. If a receiver is not designed to operate on preset frequencies selected by a pushbutton control, provision should be made for tuning to any frequency within the specified range without any impairment in gain, bandwidth, and selectivity;

— constancy of input-circuit parameters in the face of changes in antenna parameters and the input impedance of the next stage in the receiver. This is an important requirement with untuned antennas which tend to cause losses in the input circuit, a fact which might lead to an extended bandwidth, an impaired selectivity, and changes in the tuning of the input circuit.

2.2. Receiving Dummy Antennas

A receiving antenna may be represented by an equivalent generator of emf (open-circuit voltage) $\dot{E}_A$ or of current $\dot{I}_A$ (Fig. 2.5). In the general case, the impedance of the emf generator would contain a resistive and a reactive component, that is

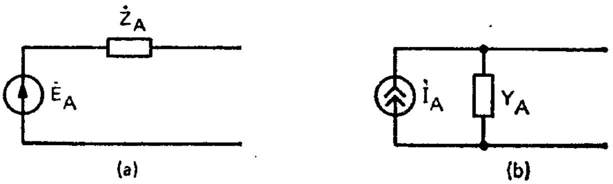


Fig. 2.5

$$Z_A = r_A + jx_A$$

(see, for example, Fig. 1.3). The generator emf is given by

$$\dot{E}_A = \dot{\varepsilon}_s h_{eff}$$

where $\dot{\varepsilon}_s$ is the electric component of the signal field at the point of reception, and h_{eff} is the effective height of the antenna.

The current generator is characterized by

$$\dot{I}_A = \dot{E}_A/\dot{Z}_A = \dot{E}_A\dot{Y}_A \qquad (2.1)$$

where

$$\dot{Y}_A = 1/\dot{Z}_A = G_A + jB_A$$

is the complex admittance of the antenna,

$$G_A = r_A/|\dot{Z}_A|^2, \qquad (2.2a)$$

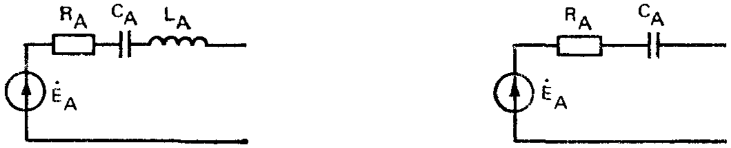
is the conductance of the antenna, and

$$B_A = -x_A/|\dot{Z}_A|^2 \qquad (2.2b)$$

is the susceptance of the antenna.

The impedance of an untuned antenna, Z_A, depends on frequency in a complex way because an antenna is a distributed-constant (or distributed-parameter) network. If an antenna is small in comparison with the wavelength in question, it may be equivalently represented by a series network of inductance L_A, capacitance C_A and resistance R_A, as shown in Fig. 2.6*a*. In the MF and LF bands, $\omega L_A \ll \ll 1/\omega C_A$, therefore the inductance may be neglected. Then the antenna equivalent circuit will only contain C_A and R_A, as shown in Fig. 2.6*b*. In the HF band, the impedance of untuned antennas may be both capacitive and inductive in its effect.

In the VHF band and at still higher frequencies, use is made of antennas which are tuned to the midband frequency where the



(a) (b)

Fig. 2.6

antennas have a resistance R_A. If it is equal to the characteristic impedance of the feeder, ρ_A, the antenna may be connected to it directly; in all other cases, this may only be done via a matching device. Then the antenna and feeder will be together equivalent to a generator of emf $E_{\acute{A}}$, having a resistance ρ_A, or to a generator of current $I_A = E_A/\rho_A$, having a conductance $G_A = 1/\rho_A$.

At microwave frequencies*, instead of emf or current, one may consider the nominal signal power in the antenna because any transformer element would change the voltage and current whereas the power would remain unchanged. The nominal power is proportional to the effective antenna area, A_{eff}:

$$P_{nom} = (\varepsilon_s^2/120\pi)\, A_{eff}\eta_A$$

where η_A is the efficiency of the antenna operating into a matched load.

2.3. Frequency Coverage

A tuned circuit can be tuned to any frequency within a specified frequency band by varying either the tuned-circuit inductance or the tuned-circuit capacitance, or both. Consider tuning by variation of the tuned-circuit inductance, with the tuned-circuit capacitance held constant. To begin with, express the tuned-circuit parameters

* The term 'microwave' applies to radio waves in the frequency range of 300 MHz and upward or wavelengths from decimetric through centimetric to millimetric and shorter.— *Translator's note.*

in terms of its capacitance

$$\rho = 1/\omega_0 C_{\text{ckt}}$$

$$d_{\text{ckt}} = r/\rho = r\omega_0 C_{\text{ckt}}$$

where ρ is the characteristic impedance and d_{ckt} is the damping factor of the tuned circuit.

Owing to skin effect in the tuned-circuit coil conductors and dielectric loss, r increases approximately in proportion to frequency. On the other hand, d_{ckt} is proportional to the frequency squared, the bandwidth is given by

$$B_{0.7} = f_0 d_{\text{ckt}}$$

and the tuned-circuit conductance at resonance

$$G_0 = d_{\text{ckt}}/\rho = \omega_0 C_{\text{ckt}} d_{\text{ckt}}$$

is proportional to the frequency cubed. Therefore, when a tuned circuit is tuned to the desired frequency by varying its inductance, its parameters markedly change over the band, which is an unwanted development.

When tuning is done by varying the tuned-circuit capacitance, its parameters may be expressed in terms of its inductance as follows:

$$\rho = \omega_0 L_{\text{ckt}}$$

$$d_{\text{ckt}} = r/\rho = r/\omega_0 L_{\text{ckt}}$$

Assuming, as before, that r is proportional to frequency, it can be seen that the damping factor and, in consequence, the Q-factor ($Q_{\text{ckt}} = 1/d_{\text{ckt}}$) of the tuned circuit are independent of frequency. The bandwidth and the equivalent resonance resistance of the tuned circuit

$$R_0 = 1/G_0 = \omega_0 L_{\text{ckt}} G_{\text{ckt}}$$

are proportional to frequency. In consequence, capacitive tuning is accompanied by less marked changes in the properties of the tuned circuit. That is the reason why, given a relatively wide frequency range, tuned circuits are tuned by varying their capacitance. Inductive tuning by, say, moving a magnetic core, or "slug", inside the tuned-circuit coil is used when the bandwidth is narrow and also when tuning by means of a variable capacitor is undesirable for constructional reasons, for example in automobile receivers which have to operate under heavy vibrations.

With capacitive tuning, the maximum-to-minimum frequency ratio over the band is

$$k_{\text{b}} = f_{\text{max}}/f_{\text{min}} = (C_{\text{max}}/C_{\text{min}})^{1/2},$$

As a rule, $k_{\text{b}} \leqslant 3$.

As has been noted (see Sec. 1.3), the frequency range covered by a receiver is divided into bands. A change-over from one band to another is done by switching tuned-circuit coils. If continuous tuning is done by varying the tuned-circuit inductance, bands are selected by switching the tuned-circuit capacitors.

If the overall frequency range is divided into equal bands (that is, those spanning equal frequency intervals, as shown in Fig. 2.7), the difference between their maximum and minimum frequencies is likewise the same

$$f_{i,\max} - f_{i,\min} = \Delta f_{\text{b}} = \text{const}$$

The maximum-to-minimum frequency ratio of the bands

$$\begin{aligned} k_{\text{b}} &= f_{i,\max}/f_{i,\min} = (f_{i,\min} + \Delta f_{\text{b}})/f_{i,\min} \\ &= 1 + \Delta f_{\text{b}}/f_{i,\min} \end{aligned}$$

decreases in going to the upper bands. To minimize the effect of the tuned-circuit capacitor C_{ckt} on the operating frequency, the tuned circuits are extended to include a padder capacitor, C_1, and a trimmer capacitor, C_2, connected as shown in Fig. 2.8. An advantage of this division into bands is that it gives the same tuning density (that is, the same number of stations per scale division) on each band.

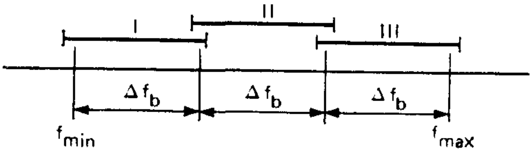


Fig. 2.7

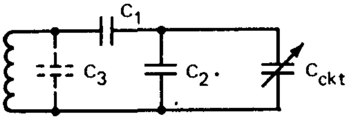


Fig. 2.8

When the division into bands is such that the maximum-to-minimum frequency ratio is the same on each band, fewer bands will be required than in the previous case, and the width of each band

$$\Delta f_{\text{b}} = f_{i,\max} - f_{i,\min} = k_{\text{b}} f_{i,\min} - f_{i,\min} = (k_{\text{b}} - 1)\, f_{i,\min}$$

will increase with increasing $f_{i,\min}$. In consequence, the tuning density will increase, too.

2.4. Electronic Tuning

For over a half-century, the principal control for adjusting the natural frequency of tuned circuits has been the tuning capacitor, that is, a variable capacitor in which the movable plates (making up the rotor) are mechanically rotated relative to the stationary plates (making up the stator). It is still being used, but in most receivers it has given way to the *varactor* (also called a *varactor diode*, a *silicon capacitor*, and a *voltage-variable capacitor*) in which the capacitance is controlled by varying the applied d.c. voltage. Among the most valuable advantages offered by varactors are simplicity of automatic and remote tuning control, small size, and mechanical reliability.

Figure 2.9 shows how a varactor can be connected into a manually adjusted tuned circuit. In each case, the control voltage is taken from a regulated power supply via a sliding-contact voltage divider. Such dividers are not reliable enough because of contact wear and low mechanical strength. Therefore, use is made of electronic sources of control voltage, such as a pulse counter energized by a pulse generator. Resistor *R* is provided in order to minimize the shunting effect of the tuning control circuit on the tuned circuit.

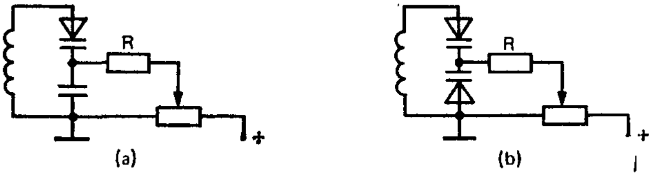


Fig. 2.9

The varactor capacitance depends not only on the d.c. control voltage, but also on the r.f. a.c. voltages existing in the tuned circuit. Therefore, the varactor may contribute to the nonlinear frequency conversion of both the desired and interfering signals, thereby impairing selectivity. The nonlinear effects may be mitigated by the same approach as permits a reduction in nonlinear distortion in amplifiers, namely by using balanced (push-pull) circuits. In a push-pull circuit, the two varactors are connected back-to-back as shown in Fig. 2.9*b*.

In the past, bands were selected by sliding-contact switches, such as shown in Fig. 2.10*a*. Unfortunately, such switches had a low reliability and were the main cause of receiver failure.

A further drawback of band selection by switches consists in that it is difficult to implement automatic and remote control. In part, this limitation can be overcome through the use of a bank of relays each of which, when its coil is energized, completes the circuit to one of the tuned-circuit coils. This purpose can best be served by what are known as *sealed-contact reed relays* or *ferreeds*. When the relay coil is energized by the control current I_C, the thin ferromagnetic reed contacts are attracted to each other and complete the external circuit (see Fig. 2.10*b*). The reeds are enclosed in an evacuated glass envelope, a fact which ensures reliability, stability, and a long service life.

Another simple and reliable alternative is offered by switching diodes used as electronic switches. Their connection in the case at hand is shown in Fig. 2.10*c*. Unfortunately, the resistance of the diode is included in the tuned circuit and increases its damping factor. Also, given strong interference, the nonlinearity of electronic switches may have an adverse effect on receiver selectivity.

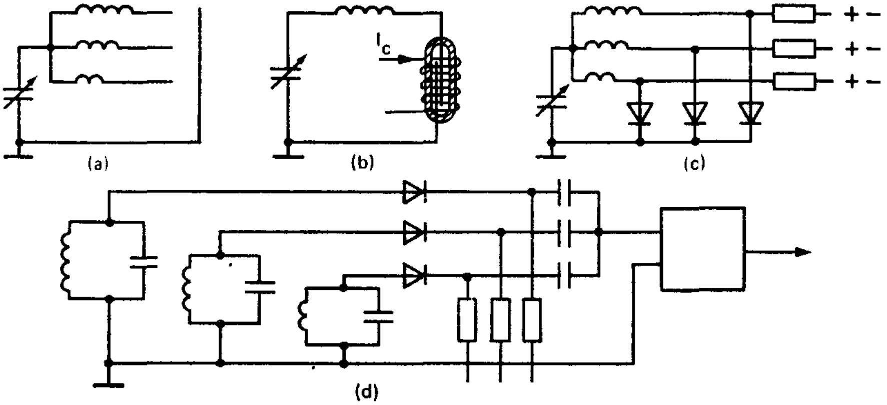


Fig. 2.10

It is seen from the foregoing that band selection is a more challenging task than electronic tuning. This is one of the reasons why receiver designers try to avoid band selection altogether. It may also be added that diode switches will affect the damping factor of tuned circuits very little if they switch complete tuned circuits, as shown in Fig. 2.10*d* rather than tuned-circuit coils.

If the first intermediate frequency of a superhet receiver has been chosen to lie above the maximum frequency of the receiver range, the suppression of spurious responses is simplified because the image and intermediate frequencies will then fall outside the receiver range. If so, one may limit oneself to the use of a preselector in the form of a broadband low-pass filter with a cutoff frequency lying above the maximum frequency of the range but below the intermediate frequency. Then there will be no need to retune the preselector and no need to use a band selector switch. Switching may be avoided in the local oscillator as well, if one uses a frequency synthesizer as the local oscillator.

The suppression of further images at the inputs of the succeeding frequency converters in such a receiver, the infradyne, is likewise simplified because these image frequencies are constant.

A broadband preselector may, however, impair multisignal selectivity, should interfering signals from nearby high-power transmitters fall within its passband. That is why it is usual to place filters at the input of professional infradyne receivers so as to suppress such interfering signals. These filters are usually tuned to a fixed frequency and have therefore a very simple design.

2.5. Analysis of a Single-Tuned Input Circuit

Single-tuned input circuits mainly differ in how the tuned circuit is coupled to the antenna and to the next receiver stage. The general relationships characterizing single-tuned input circuits at a given frequency are independent of the form of coupling.

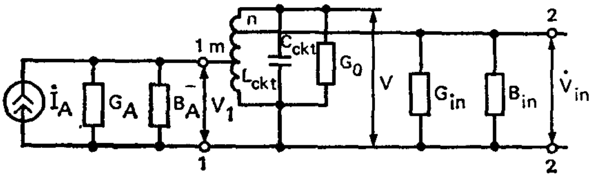


Fig. 2.11

Consider the properties of input circuits by reference to the equivalent circuit of Fig. 2.11. Here, the antenna circuit is represented by a current generator

$$\dot{I}_A = \dot{E}_A/\dot{Z}_A$$

of conductance G_A and susceptance B_A which, in the general case include the parameters of the elements that couple the antenna to the tuned circuit, namely

$$r_A = r_{ant} + r_c$$

$$x_A = x_{ant} + x_c$$

where r_{ant} = series loss resistance of the antenna proper
x_{ant} = reactance of the antenna proper
r_c = resistance of the elements coupling the antenna or feeder to the tuned circuit
x_c = reactance of the elements coupling the antenna or feeder to the tuned circuit

The input of the next stage is represented by the complex admittance

$$\dot{Y}_{in} = G_{in} + jB_{in}$$

In the diagram of Fig. 2.11, both the antenna circuit and the next stage are tapped down on the tuned-circuit coil, the respective tapping-down factors being

$$m = V_1/V \qquad (2.3)$$

$$n = V_{in}/V$$

Here m is the tapping-down factor on the antenna side, and n is tapping-down factor on the next-stage side.

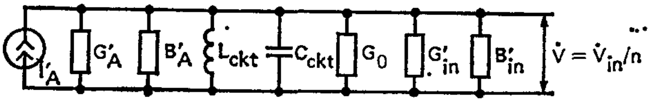


Fig. 2.12

With a tapping-down factor of less than unity, the following transformed conductances and susceptances are coupled into the tuned circuit:

$$G'_A \approx m^2 G_A \qquad (2.4)$$

$$G'_{in} \approx n^2 G_{in}$$

and

$$B'_A \approx m^2 B_A$$

$$B'_{in} \approx n^2 B_{in} \qquad (2.5)$$

Therefore, the equivalent circuit of Fig. 2.11 may be redrawn as shown in Fig. 2.12.

In view of Eq. (2.5), the equivalent susceptance of the input circuit is

$$B_{eq} = \omega C_{ckt} - 1/\omega L_{ckt} + m^2 B_A + n^2 B_{in} \qquad (2.6)$$

As is seen from Eq. (2.6), the input circuit has an equivalent capacitance C and an equivalent inductance L which are functions of the antenna and next-stage parameters transferred into the tuned circuit.

In agreement with Eq. (2.4), the equivalent conductance is

$$G_{eq} = 1/R_{eq} = G_0 + m^2 G_A + n^2 G_{in} \qquad (2.7)$$

where

$$G_0 = d_{ckt}/\rho = d_{ckt}\omega_0 C$$

is the internal loss conductance of the tuned circuit.

In view of Eqs. (2.6) and (2.7) the equivalent circuit takes the form shown in Fig. 2.13. By Ohm's law, the tuned-circuit voltage is

$$\dot V = \dot I'_A/\dot Y_{eq} = m\dot I_A/\dot Y_{eq} \qquad (2.8)$$

where $\dot{Y}_{\text{eq}}$ is the complex admittance of the equivalent circuit, defined as

$$\begin{aligned}\dot{Y}_{\text{eq}} &= G_{\text{eq}} + j\omega C + 1/j\omega L\\ &= G_{\text{eq}}\,[1 + (j\omega_0 C/G_{\text{eq}})\,(\omega/\omega_0 - \omega_0/\omega)]\\ &= G_{\text{eq}}\,(1 + j\xi)\end{aligned}$$

where $\xi = (1/d_{\text{eq}})\,(\omega/\omega_0 - \omega_0/\omega) = y/d_{\text{cq}}$ is the generalized detuning, and $y = f/f_0 - f_0/f$. At low values of ξ, $y \approx 2\Delta f/f_0$, where

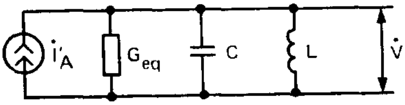


Fig. 2.13

$\Delta f = f - f_0$ and $f_0 = 1/2\pi\,(LC)^{1/2}$. The resultant damping factor of the tuned circuit is

$$d_{\text{eq}} = \rho G_{\text{eq}} = \rho\,(G_0 + m^2 G_{\text{A}} + n^2 G_{\text{in}})$$

In view of Eqs. (2.3) and (2.8), the voltage at the input of the next stage is

$$\dot{V}_{\text{in}} = n\dot{V} = mn\dot{I}_{\text{A}}/[G_{\text{cq}}\,(1 + j\xi)] = mn\dot{I}_{\text{A}}R_{\text{eq}}/(1 + j\xi) \qquad (2.9)$$

Hence, the complex gain is

$$\dot{K} = \dot{V}_{\text{in}}/\dot{E}_{\text{A}} = mnR_{\text{eq}}/[\dot{Z}_{\text{A}}\,(1 + j\xi)] \qquad (2.10)$$

The above expression gives the amplitude and phase characteristics of the input circuit.

The magnitude of the gain is

$$K = mnR_{\text{cq}}/[|\,\dot{Z}_{\text{A}}\,|\,(1 + \xi^2)^{1/2}] \qquad (2.11)$$

At resonance ($\xi = 0$),

$$K_0 = mnR_{\text{eq}}/|\,\dot{Z}_{\text{A}0}| = \frac{mn}{|\,\dot{Z}_{\text{A}0}\,|\,(G_0 + m^2 G_{\text{A}} + n^2 G_{\text{in}})} \qquad (2.12)$$

where

$$|\,\dot{Z}_{\text{A}0}\,| = (r^2_{\text{A}0} + x^2_{\text{A}0})^{1/2}$$

is the magnitude of the antenna-circuit impedance at the resonance frequency of the input tuned circuit. On the basis of Eqs. (2.11) and (2.12), the frequency response of the circuit is found to be described by the following equation:

$$\sigma = K_0/K = (|\,\dot{Z}_{\text{A}}\,|/|\,\dot{Z}_{\text{A}0}\,|)\,(1 + \xi^2)^{1/2} \qquad (2.13)$$

Far away from resonance ($\xi \gg 1$), Eq. (2.13) takes the form

$$\sigma \approx (|\,\dot{Z}_{\text{A}}\,|/|\,\dot{Z}_{\text{A}0}\,|)\,(1/d_{\text{eq}})\,(|\,f/f_0 - f_0/f\,|) \qquad (2.14)$$

At low values of ξ and neglecting the dependence of Z_A on frequency, we get

$$K_0/K = 1/\gamma = [1 + (2\Delta f/f_0 d_{eq})^2]^{1/2} \tag{2.15}$$

which checks with the equation for the frequency response of an isolated tuned circuit. Equation (2.15) can be used to derive the bandwidth of the input circuit at the specified flatness of, or variations in, the response, γ:

$$B_\gamma = f_0 d_{eq}\,(1/\gamma^2 - 1)^{1/2}$$

In the special case of $\gamma = 1/\sqrt{2} = 0.707$ (or 3 dB)

$$B_\gamma = f_0 d_{eq}$$

The phase-vs-frequency response of the input circuit is

$$-\varphi = \arctan \xi + \arctan (x_A/r_A) \tag{2.16}$$

It is seen from Eq. (2.12) that m affects the gain in a two-fold manner. If m is decreased, the numerator in (2.12) will also decrease, but at the same time the tuned circuit will be less shunted by the antenna-circuit conductance G_A, which fact is accounted for by m^2 in the denominator. A similar effect is produced by n. Let us denote

$$D = d_{eq}/d_{ckt} = G_{eq}/G_0 = \frac{G_0 + m^2 G_A + n^2 G_{in}}{G_0} \tag{2.17}$$

Then, in agreement with Eq. (2.12)

$$K_0 = mn/(|\dot{Z}_{A0}|\,DG_0) \tag{2.18}$$

Consider the condition for K_0 to be a maximum at a specified overall tuned-circuit damping factor d_{eq}, that is, let us deem that D = const. From Eq. (2.17), we get

$$m = \left[\frac{(D-1)\,G_0 - n^2 G_{in}}{G_A}\right]^{1/2} \tag{2.19}$$

and, on substituting in Eq. (2.18), we obtain

$$K_0 = \frac{n}{|\dot{Z}_{A0}|\,DG_0}\left[\frac{(D-1)\,G_0 - n^2 G_{in}}{G_A}\right]^{1/2} \tag{2.20}$$

In order to analyse Eq. (2.20) for an extremum, we equate the derivative dK_0/dn to zero and find that K_0 is a maximum at

$$n_{opt} = \left[\frac{D-1}{2}\,\frac{G_0}{G_{in}}\right]^{1/2} \tag{2.21}$$

On substituting (2.21) into (2.19), we get

$$m_{opt} = \left[\frac{D-1}{2}\,\frac{G_0}{G_A}\right]^{1/2} \tag{2.22}$$

Subject to Eqs. (2.21) and (2.22), it follows from Eq. (2.18) that the maximum gain at a specified value of d_{eq} is

$$K_{0,\max} = \frac{1}{[2\,(r_A G_{\text{in}})]^{1/2}}\,(1 - 1/D) \tag{2.23}$$

It is seen from Eqs. (2.21) and (2.22) that the gain is a maximum when the tuned circuit is equally shunted on both the antenna-circuit side and the next-stage side, that is, when

$$m^2 G_A = n^2 G_{\text{in}} = (D - 1)\,G_0/2$$

In operation with a tuned antenna, it is sought to match the antenna circuit to the receiver input. The condition of perfect match presumes the equality between the conductance coupled from the antenna into the tuned circuit, and the tuned-circuit conductance, subject to the effect of the next-stage input:

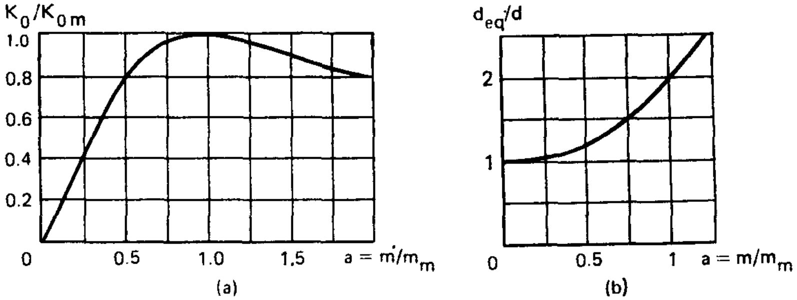


Fig. 2.14

$$m^2 G_A = G_0 + n^2 G_{\text{in}} \tag{2.24}$$

It follows from Eq. (2.24) that the tapping-down factor required for a perfect match is

$$m_{\text{m}} = [(G_0 + n^2 G_{\text{in}})/G_A]^{1/2} \tag{2.25}$$

The gain at resonance and perfect match can be found from Eq. (2.12) subject to Eqs. (2.24), (2.25) and (2.2)

$$K_{0,\text{m}} = n/2m_{\text{m}} \mid \dot{Z}_{A0} \mid G_A = n/2\,[r_A\,(G_0 + n^2 G_{\text{in}})]^{1/2} \tag{2.26}$$

For an arbitrary value of m and subject to Eqs. (2.24) through (2.26), we find from Eq. (2.12)

$$K_0 = K_{0,\text{m}} 2a/(1 + a^2) \tag{2.27}$$

where $a = m/m_{\text{m}}$ is the relative coupling coefficient. It is seen from the plot of Fig. 2.14*a* relating $K_0/K_{0\text{m}}$ to a that as the amount of coupling deviates from the optimal one by a factor of two, the gain decreases by a mere 20%.

As follows from Eq. (2.26), the gain of the input circuit at perfect match is a function of n which shows how the next stage is tapped down on the tuned-circuit coil. Find the value of n such that the resultant tuned-circuit damping factor has a specified value

$$d_{\text{eq}} = \rho G_{\text{eq}} = \rho\,(G_0 + m^2 G_{\text{A}} + n^2 C_{\text{in}}) = d\,(1 + a^2) \qquad (2.28)$$

where

$$d = \rho\,(G_0 + n^2 G_{\text{in}}) = d_{\text{ckt}} + n^2 \rho G_{\text{in}} \qquad (2.29)$$

is the tuned-circuit damping factor, considering the damping coupled in from the next-stage side.

A plot of d_{eq}/d as a function of a, defined by Eq. (2.28), is shown in Fig. 2.14*b*. As the amount of coupling between the tuned circuit and the antenna increases, the damping factor rapidly increases, and the selectivity decreases. At perfect match ($d = 1$), the resultant damping factor is

$$d_{\text{eq}} = 2d = 2\,(d_{\text{ckt}} + n^2 \rho G_{\text{in}}) \qquad (2.30)$$

Hence,

$$n_{\text{m}} = \left[\frac{d_{\text{eq}} - 2d_{\text{ckt}}}{2\rho G_{\text{in}}}\right]^{1/2} = \left[\frac{D-2}{2}\,\frac{G_0}{G_{\text{in}}}\right]^{1/2} \qquad (2.31)$$

where $D = d_{\text{eq}}/d_{\text{ckt}}$ is the coefficient of shunting, which defines the allowable increase in the resultant damping factor over d_{ckt}.

On substituting (2.31) into (2.25) and (2.26), we get

$$m_{\text{m}} = [(D/2)\,(G_0/G_{\text{A}})]^{1/2} \qquad (2.32)$$

and

$$K_{0,\text{m}} = [(1/r_{\text{A}} G_{\text{in}})\,(D - 2)/D]^{1/2}/2 \qquad (2.33)$$

It is an easy matter to see from Eq. (2.33) that the tuned circuit should be built to have the least possible own damping factor. If $D \gg 2$, then

$$K_{0,\text{m}} = (1/r_{\text{A}} G_{\text{in}})^{1/2}/2$$

In consequence, on neglecting d_{ckt}, we find from Eq. (2.30)

$$d_{\text{eq}} \approx 2n^2 \rho G_{\text{in}}$$

This is the case when the input stage is built around a FET. With a FET, it is usual that $G_0 \gg G_{\text{in}}$. In such a case, the tuned-circuit damping factor is independent of n, that is, $d_{\text{eq}} \approx 2d_{\text{ckt}}$. Therefore, one takes $n = 1$. Then, as follows from Eq. (2.26)

$$K_{0,\text{m}} = 1/2\,(r_{\text{A}} G_0)^{1/2} = (R_0/r_{\text{A}})^{1/2}/2$$

With stringent requirements for selectivity, it is advisable to reduce the amount of coupling to the antenna. As is seen from Eq. (2.28), at $a = 0.5$, the effect of the antenna causes the damping factor to increase by as little as 25%. At the same time, as follows from Eq. (2.27), the gain decreases by 20%. In operation with a tuned antenna and a long feeder line, mismatch is undesirable, as this can give rise to multiple signal reflections likely to corrupt the received message.

2.6. Input Circuits in Operation with Untuned Antennas

Untuned antennas are widely used with LF, MF and HF receivers. Since untuned antennas present a complex impedance, they couple losses and detuning into the input tuned circuit. The amount of detuning is different for different antennas, and cannot therefore be made up for completely by receiver alignment at the factory. If, however, loose coupling with the antenna is chosen such that the detuning of the tuned circuit remains within specified limits, an opportunity presents itself for operation with various antennas widely differing in parameters. A further advantage of loose coupling to the antenna is the fact that the damping coupled into the tuned circuit is not over 10% to 20% of the natural one, and this permits the selectivity of the input circuit to be preserved. Also, when loosely coupled, the input circuit has a low gain. This may be tolerated because external radio interference within the frequency bands listed substantially exceeds receiver noise.

In the receivers under consideration, the input electron device usually is a BJT or a FET. As has been noted, a FET is connected to a tap on all of the tuned-circuit coil (the tapping-down factor $n = 1$). A BJT is tapped down on the tuned-circuit coil so as to preserve the selectivity of the tuned circuit because a BJT has a low input resistance.

Continuous tuning within a band is effected by a tuning capacitor or a varactor, C_{ckt}. The overall tuned-circuit capacitance is

$$C = C_{\text{ckt}} + n^2 C_{\text{in}} + C_{\text{w}}$$

where C_{w} is the wiring capacitance.

The input-circuit gain at resonance, Eq. (2.12), is a function of the tuned-circuit impedance at resonance, R_{eq}, and admittance, $1/|\dot{Z}_{\text{A0}}|$. The antenna circuit has a resonance frequency of its own which is a function of the parameters of the antenna and of the components that provide coupling with the input tuned circuit. The plot of $1/|\dot{Z}_{\text{A0}}|$ as a function of frequency is in effect the frequency response characteristic of the antenna circuit. Variations in the input-circuit gain depend on whether the natural frequency of the antenna circuit lies above the upper or below the lower edge frequency of the band.

Consider a transformer-coupled input circuit. Referring to Fig. 2.1,

$$m = M/L_{\text{ckt}} \tag{2.34}$$

On substituting Eq. (2.34) and $R_{eq} = \omega_0 L_{ckt} Q_{eq}$ from Eq. (2.12), we find the gain at resonance

$$K_0 = \omega_0 M n Q_{eq} / |\dot{Z}_{A0}| \tag{2.35}$$

If we neglect losses in the antenna circuit in comparison with the reactance, then

$$|\dot{Z}_A| \approx x_A = |\omega L_A - 1/\omega C_A| = \omega L_A |1 - \omega_A^2/\omega^2|$$
$$|\dot{Z}_{A0}| \approx x_{A0} = \omega_0 L_A |1 - \omega_A^2/\omega_0^2| \tag{2.36}$$

where $L_A = L_{ant} + L_c$, and $\omega_A = 1/(L_A C_A)^{1/2}$ is the natural angular frequency of the antenna circuit.

Substitution of (2.36) into (2.35) gives

$$K_0 = \frac{nMQ_{eq}}{L_A |1 - \omega_A^2/\omega_0^2|} = \frac{k (L_{ckt}/L_A)^{1/2}}{|1 - \omega_A^2/\omega_0^2|} nQ_{eq} \tag{2.37}$$

where $k = M/(L_A L_{ckt})^{1/2}$.

It is seen from Eq. (2.37) that variations in the gain are different, depending on ω_A/ω_0. Consider some of the likely cases.

1. The natural frequency of the antenna circuit exceeds the upper band-edge frequency, $f_A > f_{max}$ (Fig. 2.15*a*). In the circumstances, the gain abruptly rises with increasing frequency. The point is that an increase in frequency leads to a simultaneous increase in $R_{eq} = \omega_0 L_{ckt} Q_{eq}$ and $1/|\dot{Z}_{A0}|$ because the tuning frequency of the input tuned circuit moves closer to the resonance frequency of the antenna circuit. When $f_A^2 \gg f_{max}^2$, we obtain from Eq. (2.37)

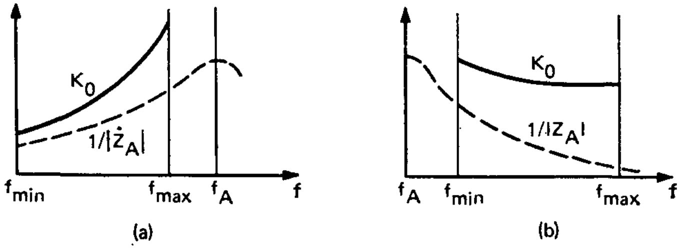


Fig. 2.15

$$K_0 \approx k (L_{ckt}/L_A)^{1/2} (\omega_0^2/\omega_A^2) nQ_{eq} \tag{2.38}$$

If $n =$ const and $Q_{eq} =$ const, then

$$K_0 = \omega_0^2 \times \text{const} \tag{2.39}$$

The flatness of the gain over the band is given by

$$F_k = K_{0,max}/K_{0,min} = \omega_{0,max}^2/\omega_{0,min}^2 = k_b^2 \tag{2.40}$$

2. The natural frequency of the antenna circuit lies below the lower band-edge frequency, $f_A < f_{min}$ (Fig. 2.15*b*). Now the gain does not vary so abruptly as in the previous case. The point is that as the antenna circuit moves away from its natural frequency, $1/|\dot{Z}_{A0}|$ decreases and R_{eq} increases, thus making up in part for the decrease in $1/|Z_{A0}|$.

When $f_A^2 \ll f_{min}^2$, it follows from Eq. (2.37) that

$$K_0 \approx k\,(L_{ckt}/L_A)^{1/2}\,nQ_{eq} \tag{2.41}$$

If $n = \text{const}$ and $Q_{eq} = \text{const}$, then

$$K_0 \approx \text{const} \tag{2.42}$$

The conditions in which Eqs. (2.39) and (2.42) have been derived are typical of receivers using FETs. With a BJT, the value of Q_{eq} is frequency-dependent owing to the coupled-in loss, $n^2\rho G_{in}$. If n is independent of frequency, then Q_{eq} falls off with increasing frequency. Therefore, K_0 in Eq. (2.41) decreases with increasing frequency.

When the tuned circuit is coupled to the amplifier by internal capacitive coupling, as shown in Fig. 2.16,

$$n = C/C_1 = 1/\omega_0^2 L_{ckt} C_1 = \text{const}/\omega_0^2 \tag{2.43}$$

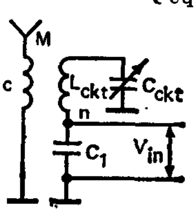


Fig. 2.16

On substituting (2.43) in (2.38), it is seen that K_0 is independent of ω_0 and proportional to Q_{eq}. If we neglect the loss coupled from the antenna circuit, then

$$d_{eq} = d_{ckt} + n^2\rho G_{in} = d_{ckt} + G_{in}/\omega_0^3 L_{ckt} C_1^2 \tag{2.44}$$

As the frequency increases, the Q-factor increases too. This serves to maintain the selectivity over the band at a constant value.

Far away from resonance, the selectivity can be found from Eq. (2.14) subject to Eq. (2.36)

$$\sigma = |1 - \omega_A^2/\omega^2|\,|\omega^2/\omega_0^2 - 1|/d_{eq}\,|1 - \omega_A^2/\omega_0^2|$$

Close to resonance, use may be made of Eq. (2.15).

3. The natural frequency of the antenna circuit lies within the operating frequency range of the receiver, $f_{min} < f_A < f_{max}$. Now the gain depends on frequency so heavily that this mode of operation is not ordinarily used.

In the case of external capacitive coupling, the input tuned circuit is coupled to the antenna via a capacitor, C_b, as shown in Fig. 2.2. In order that the antenna parameters could affect the tuning of the tuned circuit as little as possible, the capacitance of C_b should be low. In consequence, the total capacitance of the series network of C_{ant} and C_{b} should likewise be low. Let us denote

$$C_{\mathrm{A}} = C_{\mathrm{ant}}C_{\mathrm{b}}/(C_{\mathrm{ant}} + C_{\mathrm{b}})$$

The reactance $1/\omega C_{\mathrm{A}}$ is many times ωL_{A} and r_{A}, so they may be neglected. Under these assumptions,

$$|\dot{Z}_{\mathrm{A0}}| \approx 1/\omega_0 C_{\mathrm{A}} \qquad (2.45)$$

On substituting $R_{\mathrm{eq}} = \omega_0 L_{\mathrm{ckt}} Q_{\mathrm{eq}}$, $m = 1$, and $|\dot{Z}_{\mathrm{A0}}|$ from Eq. (2.45) in Eq. (2.12), we get

$$K_0 = n\omega_0^2 L_{\mathrm{ckt}} C_{\mathrm{A}} Q_{\mathrm{eq}} \qquad (2.46)$$

If $n = \mathrm{const}$ and $Q_{\mathrm{eq}} = \mathrm{const}$, then

$$K_0 = \omega_0^2 \times \mathrm{const} \qquad (2.47)$$

Variations in gain over the band can be found by Eq. (2.40). The quadratic relationship in Eqs. (2.46) and (2.47) is explained by the fact that an increase in frequency leads to a simultaneous increase in $1/|Z_{\mathrm{A0}}| = \omega_0 C_{\mathrm{A}}$ and R_{eq}.

If the tuned circuit is coupled to the next stage by internal capacitive coupling, as shown in Fig. 2.17*a*, then, in agreement with Eq. (2.43), we get

$$K_0 = (C_{\mathrm{A}}/C_1)\, Q_{\mathrm{eq}}$$

where $Q_{\mathrm{eq}} = 1/d_{\mathrm{eq}}$ is defined by Eq. (2.44).

If the input tuned circuit is coupled to the antenna by internal capacitive coupling, as shown in Fig. 2.17*b*, then the antenna and

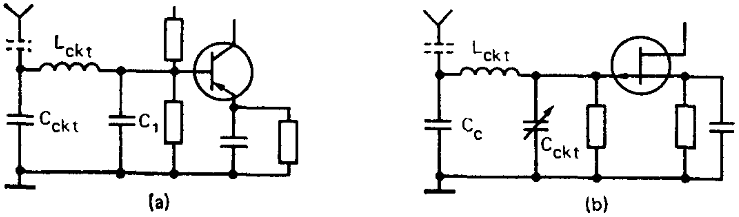


Fig. 2.17

the input of the next stage are coupled to the tuned circuit via a divider made up of C_{c} and $C_{\Sigma} = C_{\mathrm{ckt}} + C_{\mathrm{in}}$. With this arrangement, C_{c} is chosen to be many times C_{Σ} so as to provide loose coupling with the antenna. The resultant tuned-circuit capacitance is

$$C = C_{\Sigma}C_{\mathrm{c}}/(C_{\Sigma} + C_{\mathrm{c}})$$

The respective tapping-down factors are

$$\left.\begin{aligned} n &= C/C_{\mathrm{c}} = C_{\Sigma}/(C_{\Sigma} + C_{\mathrm{c}}) \approx C_{\Sigma}/C_{\mathrm{c}} \\ n &= C/C_{\Sigma} = 1/\omega_0^2 L_{\mathrm{ckt}} C_{\Sigma} \end{aligned}\right\} \qquad (2.48)$$

On neglecting the antenna resistance, we get

$$|\dot{Z}_{A0}| \approx \omega_0 L_A - 1/\omega_0 C_A = (1/\omega_0 C_A)(1 - \omega_0^2/\omega_A^2) \qquad (2.49)$$

where $\omega_A = 1/(L_A C_A)^{1/2}$ is the natural angular frequency of the antenna circuit.

Subject to Eqs. (2.48) and (2.49), we get from Eq. (2.12)

$$K_0 = (C_A/C_c)\, Q_{eq}/(1 - \omega_0^2/\omega_A^2)$$

If $\omega_A^2 \gg \omega_0^2$, then

$$K_0 = (C_A/C_c)\, Q)_{eq}$$

If Q_{eq} = const, the gain at resonance K_0 is frequency-independent, which means that the inequality $\omega_A^2 \gg \omega_0^2$ will be satisfied if the antenna is small in size or if C_b is low. The circuit configuration in Fig. 2.17*a* should preferably be used when the next stage has a low input resistance and $n \ll 1$. The configuration in Fig. 2.17*b* should be used when R_{in} is high and $n \approx 1$ may be tolerated.

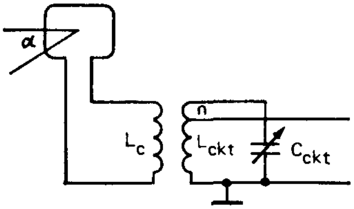


Fig. 2.18

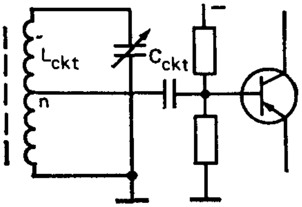


Fig. 2.19

Selectivity far away from resonance can be found by Eq. (2.14) subject to Eqs. (2.48) and (2.49)

$$\sigma = (f/f_0)(1/d_{eq})(f/f_0 - f_0/f)$$
$$= Q_{eq}(f^2/f_0^2 - 1)$$

With a loop antenna (Fig. 2.18), the antenna-circuit emf E_A is a function of the angle α between the loop plane and the direction of signal arrival

$$E_A = E_{A0} \cos \alpha$$

where $E_{A0} = \varepsilon_s h_{eff}$ is the emf of the signal arriving in the plane of the loop. The effective height of the antenna depends on the surface area of the loop, A_l, and the number of turns, N_t:

$$h_{eff} = 2\pi A_l N_t/\lambda$$

where λ is the wavelength.

The transmission gain of the input circuit with a loop antenna is given by Eq. (2.35) where

$$|\dot{Z}_{A0}| = r_A^2 + (\omega L_\Sigma)^2$$
$$L_\Sigma = L_1 + L_c$$

The size of a loop antenna can be reduced through the use of a ferrite core which enhances the signal emf by concentrating the magnetic flux. The effective height of a ferrite-rod antenna is

$$h_{\mathrm{eff}} = 2\pi A_{1} N_{\mathrm{t}} \mu_{\mathrm{c}} \psi / \lambda$$

where μ_{c} is the permeability of the ferrite core and ψ is a coefficient which takes care of the shape of the coil and its position on the core.

The antenna coil is used as the input tuned-circuit inductor, as shown in Fig. 2.19. The transmission gain of the input circuit may be found from Eq. (2.12) on substituting $m = 1$ and $|\dot{Z}_{\mathrm{A0}}| \approx \approx \omega_0 L_{\mathrm{ckt}}$:

$$K_0 = nQ_{\mathrm{eq}}$$

In receivers with a ferrite-rod antenna, the sensitivity is customarily expressed in units of signal field intensity (microvolts per metre), namely

$$\varepsilon_{\mathrm{s}} = E_{\mathrm{A}} / h_{\mathrm{eff}}$$

Then, the transmission gain in terms of signal field intensity is

$$K_{0\varepsilon} = V_{\mathrm{in}} / \varepsilon_{\mathrm{s}} = nQ_{\mathrm{eq}} h_{\mathrm{eff}}$$

Where it is essential to provide high selectivity and, at the same time, a flat gain over the band, a passband filter is ordinarily placed in the input circuit. In LF and MF broadcast receivers, this often is a two-section filter.

As the tuning of the filter is varied, this causes the gain and bandwidth to change. To avoid abrupt changes in these two variables, the tuned circuits are coupled in such a way that an increase in frequency brings about a decrease in the coupling coefficient with the result that the bandwidth remains practically constant. One way to achieve this is to combine two or more forms of coupling: internal and external capacitive or internal capacitive and transformer. In the circuit of Fig. 2.4, internal capacitive coupling is provided by capacitor C_{c1}, and external capacitive coupling by capacitor C_{c2}. The first tuned circuit in the filter is coupled to the antenna in the same way as this is done in a single-tuned circuit.

In order to find the gain at resonance for the input circuit, Q_{eq} in Eqs. (2.12), (2.37) and (2.46) should be replaced by the filter gain at resonance, $K_{0,\mathrm{f}}$. Thus, for a double-tuned input circuit, we get from Eq. (2.12)

$$K_0 = (mnR_{0,\mathrm{eq}} / |\dot{Z}_{\mathrm{A0}}|)\,[\beta/(1+\beta^2)]$$

For the circuit of Fig. 2.4

$$K_0 = k\,(L_{\mathrm{ckt}}/L_{\mathrm{A}})^{1/2}\,[nQ_{\mathrm{eq}}/(1-\omega_{\mathrm{A}}^2/\omega_0^2)]\,[\beta/(1+\beta^2)]$$

where

$$\beta = k_{\mathrm{c}} / d_{\mathrm{eq}}$$

is the normalized coupling coefficient of the tuned circuits in the filter. It varies with tuning frequency and affects changes in the gain of the input circuit over the band.

2.7. Input Circuits in Operation with Tuned Antennas

Tuned antennas are used at microwave frequencies where the antenna size is comparable with the wavelength. and also in professional HF receivers, such as are used on point-to-point radio links. In all of these applications, stringent requirements are specified for receiver sensitivity. Since sensitivity is limited by receiver noise, it is important to provide for the best possible transmission of the received signal from the antenna to the r.f. amplifier. The transmission gain is a maximum when the antenna is perfectly matched to the feeder and the feeder is perfectly matched to the receiver input. Then a travelling wave will be sustained on the feeder, which is likewise important for avoiding signal distortion caused by reflections when a long feeder is used. A typical equivalent circuit of the input circuit operating with a matched feeder line is shown in Fig. 2.11. Proper match of the feeder to the receiver input and the specified resultant damping factor are obtained by choosing appropriate values for n_{match} and m_{match} by Eqs. (2.31) and (2.32). The gain at resonance and under conditions of perfect match is given by Eq. (2.33); when proper match is not obtained, it is given by Eq. (2.27). The resonance properties of the input circuit are described by Eqs. (2.13) through (2.15).

Tuned antennas has a large bandwidth, therefore variations in the antenna impedance away from resonance need not be considered.

The mode of operation of the input circuits in question can conveniently be described in terms of the available power utilization factor

$$U_{\mathrm{P}} = P/P_{\mathrm{avail}} \qquad (2.50)$$

where

$$P = V_{\mathrm{in}}^2 G_{\mathrm{in}} \qquad (2.51)$$

is the power actually developed at the input to the next stage, and

$$P_{\mathrm{avail}} = E_{\mathrm{A}}^2/4r_{\mathrm{A}} \qquad (2.52)$$

is the available power of the antenna-feeder system.

On substituting (2.51) and (2.52) into (2.50), we obtain a simple relationship between the available-power utilization factor and the voltage gain

$$U_{\mathrm{P}} = 4K_0^2 G_{\mathrm{in}} r_{\mathrm{A}} \qquad (2.53)$$

U_P is a maximum when K_0 is a maximum. On the basis of Eqs. (2.53), (2.27) and (2.33), we get

$$U_P = \frac{D-2}{D}\left(\frac{2a}{1+a^2}\right)^2 \qquad (2.54)$$

It is seen from the foregoing that the available-power utilization factor describes the degree of mismatch between the feeder and the receiver input and the losses suffered in the input circuit. When the input tuned circuit is lossless ($d_{ckt} \approx 0$ or $D \gg 2$) and at complete match ($a = 1$), $U_P = 1$. In all other cases, $U_P < 1$.

At $a=1$, the available-power utilization factor, $U_P = (D - 2)/D$, defines the losses in the input circuit. The value of $D = d_{eq}/d_{ckt}$ is chosen according to the requirements for the selectivity and bandwidth of the input circuit. At $D \gg 2$, the available-power utilization factor, $U_P = [2a/(1 + a^2)]^2$, gives the degree of mismatch.

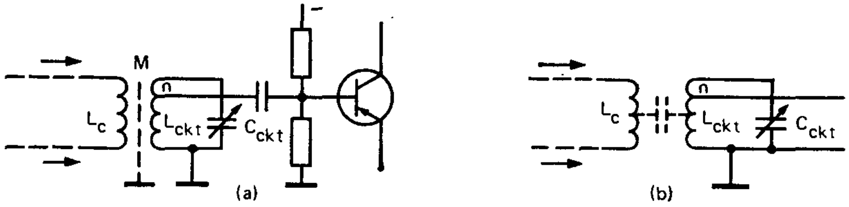


Fig. 2.20

The feeder can be coupled to the receiver input in anyone of several ways: by a tapped coil, a transformer, or a capacitive divider. They are practically of equal value if a shielded feeder is used.

Tapped-coil coupling (see Fig. 2.3) is used when the feeder is a length of coaxial cable. The desired condition of match is obtained through the choice of the tapping-down factor

$$m = (L_1 + M_1)/L_{ckt}$$

where L_1 is the inductance of the tuned-circuit coil's part between the points at which the antenna feeder is connected, and M_1 is the mutual inductance between the points of connection of the feeder and all of the tuned-circuit coil.

The equivalent circuit for this arrangement is the same as shown in Fig. 2.11. The conclusions drawn in Sec. 2.4 fully apply in this case, too, considering that $r_A = \rho_A$, $x_A = 0$, $|\dot{Z}_{A0}| = \rho_A$, and $G_A = 1/\rho_A$.

Transformer coupling (see Fig. 2.20) can be used with both a balanced and an unbalanced feeder. In the former case, transformer coupling enables the receiver to have a balanced input, which is essential in order to prevent an unshielded feeder from acting as an antenna. To this end, one uses an electrostatic shield between the coupling coil and the tuned-circuit coil, as shown in Fig. 2.20*a*, and a suitably designed feeder. Owing to the shield, coupling between the coils is solely provided by the mutual inductance M. The currents induced by the electromagnetic field directly in the feeder conductors have their paths completed in the coupling coil and cancel out. Without an electrostatic shield the capacitance between the coupling coil and the tuned-circuit coil (see Fig. 2.20*b*) would prevent the induced currents from cancelling each other, and the feeder would act as an antenna.

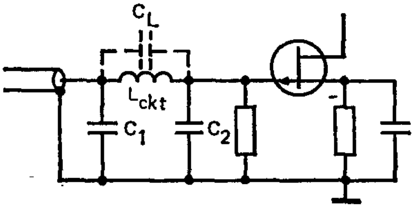


Fig. 2.21

The above configuration differs from the previous one in how the tuned circuit is coupled to the feeder. The transformation ratio (or the tapping-down factor) is given by

$$m = M/L_{\rm ckt} = k\,(L_{\rm c}/L_{\rm ckt})^{1/2}$$

where

$$k = M/(L_{\rm ckt}/L_{\rm c})^{1/2}$$

is the coupling factor. In order to find the value of the coupling factor $k_{\rm m}$ required for perfect match, resolve Eq. (2.55) for k and substitute for $m_{\rm m}$ from Eq. (2.25)

$$k_{\rm m} = m_{\rm m}\,(L_{\rm ckt}/L_{\rm c})^{1/2} = (L_{\rm ckt}/L_{\rm c})^{1/2}\,[(G_0 + n^2G_{\rm in})/G_{\rm A}]^{1/2} \qquad (2.56)$$

As follows from Eq. (2.29),

$$G_0 + n^2G_{\rm in} = d/\rho = d/\omega_0L_{\rm ckt} \qquad (2.57)$$

and, as follows from Eq. (2.2),

$$G_{\rm A} = r_{\rm A}/|\,Z_{\rm A0}\,|^2 = \rho_{\rm A}/[\rho_{\rm A}^2 + (\omega_0L_{\rm c})^2] \qquad (2.58)$$

In view of Eqs. (2.57) and (2.58), Eq. (2.56) takes the form

$$k_{\rm m} = d^{1/2}\,(\rho_{\rm A}/\omega_0L_{\rm c} + \omega_0L_{\rm c}/\rho_{\rm A})^{1/2} \qquad (2.59)$$

The physically feasible value of $k_{\rm m}$ does not exceed 0.5 or 0.6. Therefore $L_{\rm c}$ should be chosen such that perfect match is obtained at the least possible value of $k_{\rm m}$. The condition for $k_{\rm m}$ to be a minimum can be determined by solving the equation

$$dk_{\rm m}/dL_{\rm c} = 0$$

As a result, we obtain

$$k_{\rm m,min} = (2d)^{1/2} = d_{\rm eq}^{1/2} \qquad (2.61)$$

As a rule, perfect match is sought at the midband frequency. At the band edges, the amount of coupling differs only slightly from the optimum one, and the gain is a maximum very nearly.

Coupling by a capacitive divider, as shown in Fig. 2.21, is used with an unbalanced feeder. The tuned circuit is formed by the tuned-circuit inductance L_{ckt} and the capacitance

$$C = C_1 C_{2\Sigma}/(C_1 + C_{2\Sigma}) + C_{\text{L}}$$

where $C_{2\Sigma} = C_2 + C_{\text{in}}$ and C_{L} is the interturn capacitance of the tuned-circuit coil L_{ckt}.

In this circuit, where C_1 and C_2 are connected in series, the resultant capacitance is lower than it is in tuned circuits where the capacitances are connected in parallel and added together. The respective tapping-down factors are

$$m = C/C_1 \approx C_{2\Sigma}/(C_1 + C_{2\Sigma}) < 1$$

and

$$n = C/C_{2\Sigma} \approx C_1/(C_1 + C_{2\Sigma}) < 1$$

Furthermore,

$$m + n = 1$$

Therefore, if m has been chosen so as to perfect match, then

$$n = 1 - m_{\text{match}}$$

The above arrangement is advantageous in that it can be used at higher frequencies owing to a decrease in the capacitance of the tuned-circuit capacitor C.

2.8. Input Circuits for Various Frequency Bands

Lumped-parameter (or lumped-constant) tuned circuits are used in the VHF band and at lower frequencies and, in part, in the UHF band. With specified values of L and C, the resonance frequency of a tuned circuit is

$$f_0 = 1/2\pi\,(LC)^{1/2}$$

If L and C are changed by ΔL and ΔC, respectively,

$$f_0' = f_0\,[1 + \Delta L/L + \Delta C/C + (\Delta L/L)\,(\Delta C/C)]^{-1/2} \qquad (2.62)$$

At small fractional changes $\Delta L/L$ and $\Delta C/C$, if we expand Eq. (2.62) into a power series and limit ourselves to terms of the first order of smallness, we get

$$f_0' \approx f_0\,(1 - \Delta L/2L - \Delta C/2C) \qquad (2.63)$$

The departure of the resonance frequency from its initial value is

$$\Delta f = f_0 - f_0'$$

Therefore, Eq. (2.63) can be used to determine the tuning stability

$$\Delta f/f_0 \approx \Delta L/2L + \Delta C/2C \qquad (2.64)$$

It is seen from the foregoing that the decrease in L and C required for an increase in the resonance frequency has a detrimental effect on the frequency stability.

In the UHF band and the higher frequencies, use is made of resonators made from lengths of coaxial or strip lines. Among their advantages are high Q, high stability, and ruggedness.

As a rule, the loss coupled into a resonant (or tuned) line is many times its internal loss. For this reason, such a line may be taken as an ideal, lossless one. As the theory of electric circuits tells us, the input admittance of a short-circuited lossless line is

$$\dot{Y}_{\text{line}} = (-j/\rho_{\text{line}}) \cot 2\pi\,(l/\lambda)$$

where l = line length
λ = wavelength
ρ_{line} = characteristic impedance of the line

A quarter-wave short-circuited line, that is, one with $l = \lambda/4$, is not unlike a parallel resonant circuit. If the free end of the line is connected to a circuit whose input admittance is

$$Y_{\text{in}} = G_{\text{in}} + j\omega C_{\text{in}}$$

then for resonance to occur at a specified frequency one should satisfy the following condition:

$$(1/\rho_{\text{line}}) \cot 2\pi\,(l/\lambda_0) = \omega_0 C_{\text{in}}$$

Hence the required length of the resonant line is

$$l_{\text{res}} = (\lambda_0/2\pi) \text{ arc tan } (1/\rho_{\text{line}}\omega_0 C_{\text{in}})$$

Within a relatively narrow frequency range (comprising the passband and adjacent frequency bands), a resonant line may be replaced by an equivalent tuned circuit. For this purpose, we require the equality of the derivatives

$$(\partial\dot{Y}_{\text{line}}/\partial\omega)_{\omega=\omega_0} = (\partial\dot{Y}_{\text{ckt}}/\partial\omega)_{\omega=\omega_0} \qquad (2.65)$$

where

$$\dot{Y}_{\text{line}} = G_{\text{in}} + j\omega C_{\text{in}} - (j/\rho) \cot (2\pi l_{\text{ckt}}/\lambda)$$

and

$$\dot{Y}_{\text{ckt}} = G_{\text{in}} + j\omega C - j/\omega L$$

is the admittance of the tuned circuit equivalent to the resonant line.

As follows from Eq. (2.65), the capacitance of the equivalent tuned circuit is

$$C = \left[C_{\text{in}} + \frac{2\pi l_{\text{res}}}{\lambda_0\omega_0\rho_{\text{line}} \sin^2 (2\pi l_{\text{res}}/\lambda_0)}\right] \Big/ 2$$

The amount of damping coupled into the line from the next stage at a tapping-down factor of unity is

$$d = G_{\text{in}}/\omega_0 C$$

A resonant line may be coupled to an antenna feeder by autotransformer coupling as in Fig. 2.22*a*, by a transformer as in Fig. 2.22*b*, or by a capacitance as in Fig. 2.22*c*.

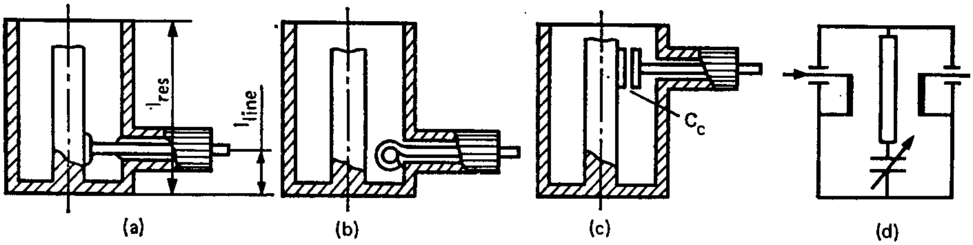


Fig. 2.22

Input circuits using strip lines are arranged in a similar manner (Fig. 2.22*d*).

Resonant lines can be tuned by varying their capacitance or their effective length l_{res}. The characteristics derived in Secs. 2.4 and 2.6 apply to line sections as well.

In the SHF and EHF bands, the input circuit usually consists of a waveguide system in which the individual sections act as what are

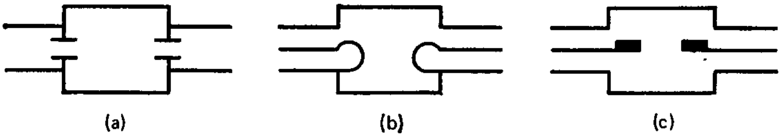


Fig. 2.23

known as *cavity resonators* or *resonant cavities*. A resonant cavity is a metal enclosure inside which a field is excited by means of a hole in the wall, as in Fig. 2.23*a*, by a loop as in Fig. 2.23*b*, or by a probe as in Fig. 2.23*c*. Among the advantages offered by cavity resonators are high Q, high stability, convenient size, and a nearly ideal shielding.

A cavity resonator can act as an impedance-transforming element. In such a case, as for strip-line resonators widely used in the SHF band, it can be represented by the equivalent circuit of Fig. 2.11 and described by the quantitative relations derived in Secs. 2.4 and 2.6.

Chapter Three

Radio Signal Amplifiers

3.1. Purpose and Principal Characteristics

In a receiver, radio signals, that is, modulated carrier waves, are amplified at r.f. prior to frequency conversion and at the i.f. after frequency conversion. The input stages should have a low noise factor, a high input impedance, and a flat gain over the band. For this reason, they are often built around FETs, BJTs, and tunnel diodes. There may be parametric and maser amplifiers.

Apart from amplification, both r.f. and i.f. amplifiers provide frequency selectivity. For this purpose, they include tuned circuits, filters made up of coupled tuned circuits, piezoceramic resonators, active filters, and other resonant circuits. Amplifiers whose frequency response is made, owing to filters, a nearly square one are referred to as bandpass amplifiers. Tunable radio-frequency (r.f.) amplifiers are most often of the single-tuned type.

The principal parameters and properties of amplifiers are the gain at resonance, bandwidth, selectivity, noise factor, signal distortion, and stability, that is, the ability of an amplifier to preserve its performance in service.

3.2. Tuned Amplifiers Using Nonreciprocal Elements

In radio-signal amplifiers, the amplifying device may be connected in any one of two ways, namely in a common-emitter or a common-base circuit in the case of BJTs; in a common-source or a common-gate configuration in the case of FETs and in a common-cathode or a common-grid circuit in the case of vacuum tubes.

When used in the HF and lower-frequency bands, common-emitter (common-source or common-cathode) amplifiers provide the highest power gain of all. On the other hand, common-base (common-gate or common-grid) amplifiers are highly stable against the effect of parasitic feedback, for which reason they are often used in the UHF and SHF bands. The arrangement and analysis of tuned amplifiers are the same for all types of amplifying devices and their circuit configurations.

Figure 3.1 shows the circuit of a common-source FET amplifier. The drain lead contains a tuned circuit, L_{ckt}, C_{ckt}. The tuned circuit is adjusted with the aid of its capacitor, C_{ckt}. The drain is series-fed via a filter, R_3C_3, and the tuned-circuit coil. Since C_3 is 50 to 100 times the maximum capacitance of the tuned-circuit capacitor, C_{ckt}, the resonance frequency is decided by L_{ckt} and C_{ckt}. At this frequency, the amplifier gain is a maximum. Its rolloff away from resonance determines the selectivity of the amplifier.

The bias voltage applied to the gate is set by the voltage drop across R_2 traversed by the source current. Capacitor C_2 eliminates a.c. negative feedback. C_1 is a d.c. blocking capacitor, and R_1 serves to feed the bias voltage to the gate.

In Fig. 3.1, the FET source is tapped down on the tuned-circuit coil—this is done in order to enhance amplifier stability. In BJT stages, the amplifying device is tapped down on the tuned-circuit coil not only to improve stability, but also to minimize the shunting of the tuned circuit by the relatively low input and output resistances. As another example, Fig. 3.2 shows an amplifier in which the first stage is tapped down on the tuned-circuit coil at m and the next stage, at n. Supply voltage to the collector is fed via a filter, R_4C_2, and some of the turns of coil L_1. The necessary d.c. values and temperature stabilization are provided by resistors R_1, R_2, and R_3. Capacitor C_1 eliminates a.c. negative feedback, and C_3 is a d.c. blocking capacitor which prevents the collector supply voltage from finding its way into the base circuit.

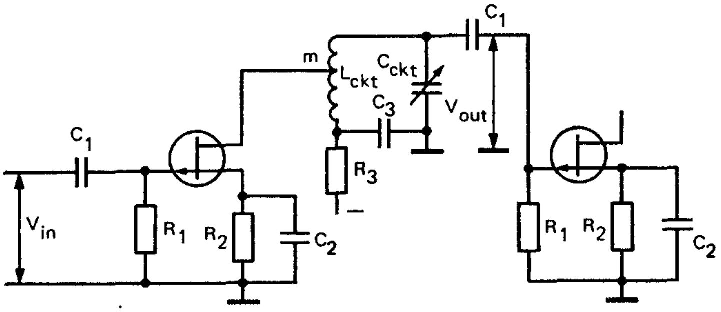


Fig. 3.1

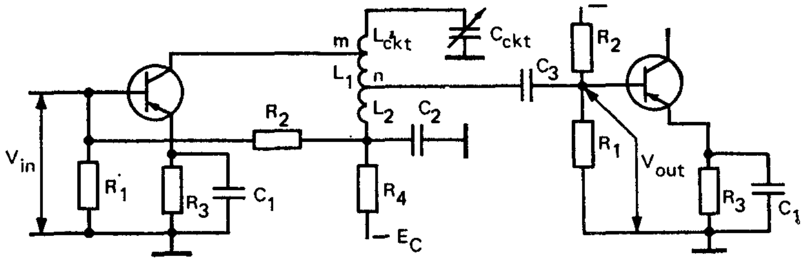


Fig. 3.2

In the circuit of Fig. 3.3, the tuned circuit is transformer-coupled to the collector of the BJT in that stage and tapped-coil-coupled to the input of the next stage.

Instead of one transistor or FET, a tuned amplifier may use a chain of two, three and more devices. Examples are the cascode amplifier and the complementary-transistor amplifier.

Intermediate-frequency (i.f.) amplifiers provide the major portion of amplification and adjacent-channel attenuation or selectance against the first channel (about 10 kHz from resonance). An important feature about them is that they are tuned to a fixed frequency. The gain of an i.f. amplifier is, as a rule, of the order of 10^4 to 10^6, which is a high figure, indeed.

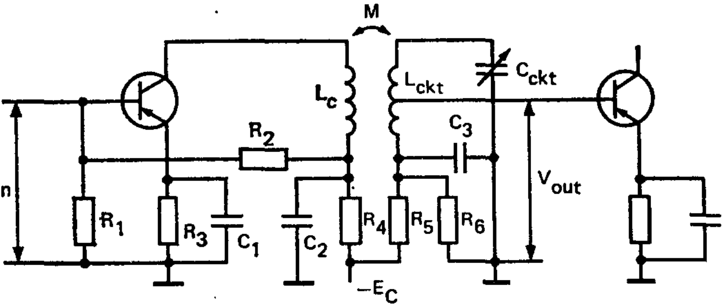


Fig. 3.3

3.3. General Analysis of the Tuned Amplifier

In small-signal analysis, an amplifying device (a BJT, a FET, an IC, or a tube) may be represented by an active linear two-port of the type shown in Fig. 3.4. In terms of the Y-parameters, the equations describing this two-port have the form

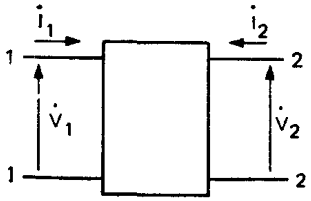


Fig. 3.4

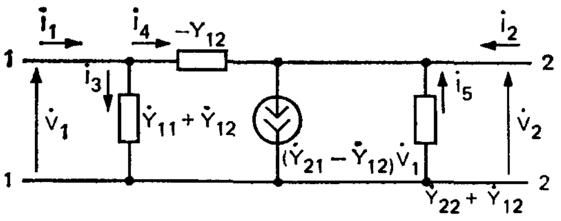


Fig. 3.5

$$\left.\begin{aligned}\dot{I}_1&=\dot{Y}_{11}\dot{V}_1+\dot{Y}_{12}\dot{V}_2\\ \dot{I}_2&=\dot{Y}_{21}\dot{V}_1+\dot{Y}_{22}\dot{V}_2\end{aligned}\right\}\qquad(3.1)$$

A widely accepted model of an amplifying device is shown in Fig. 3.5.

For further analysis, it is important to consider the dependence of the parameters on frequency. Within a limited frequency band, the Y-parameters may be written as

$$\left.\begin{aligned} \dot{Y}_{11} &\approx G_{11} + j\omega C_{11} \\ -\dot{Y}_{12} &\approx G_{12} + j\omega C_{12} \\ \dot{Y}_{21} &\approx |\dot{Y}_{21}| \exp(j\varphi_{21}) \\ \dot{Y}_{22} &\approx G_{22} + j\omega C_{22} \end{aligned}\right\} \tag{3.2}$$

where $|\dot{Y}_{21}| = g_m/[1 + (\omega/\omega_{gm})^2]^{1/2}$ (3.3)

ω_{gm} = angular frequency at which the transconductance decreases by a factor of $1/\sqrt{2}$

$$\varphi_{21} = -\text{arc tan}\,(\omega/\omega_{gm}) = -\text{arc tan}\,\omega\tau_{21} \tag{3.4}$$

and $\tau_{21} = 1/\omega_{gm}$

A complete equivalent circuit of an amplifying device includes a signal source and a load, as shown in Fig. 3.6, such that

$$\dot{I}_1 = \dot{I}_{ss} - \dot{V}_1\dot{Y}_{ss} \tag{3.5}$$

$$\dot{I}_2 = -\dot{V}_2\dot{Y}'_\Sigma \tag{3.6}$$

where

$$\dot{Y}'_\Sigma = \dot{Y}_\Sigma/m^2 = (\dot{Y}_{\text{ckt}} + n^2\dot{Y}_{\text{L}})/m^2 \tag{3.7}$$

is the total admittance of the tuned circuit and load transferred to the output of the two-port. In Eq. (3.6), the '—' sign indicates that the voltage drop across the load at terminals *2-2* due to current I_2 is opposite to V_2. The tapping-down factors are

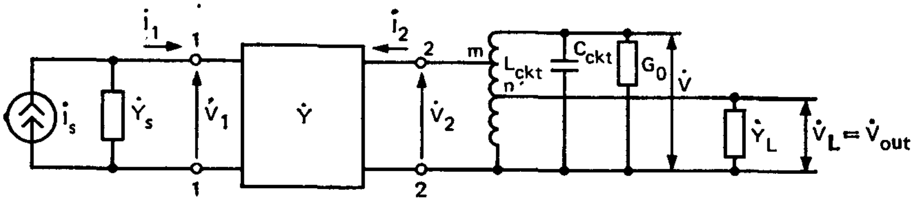


Fig. 3.6

and

$$\left.\begin{aligned} m &= \dot{V}_2/\dot{V} \\ n &= \dot{V}_{\text{L}}/\dot{V} \end{aligned}\right\} \tag{3.8}$$

Subject to Eq. (3.8), the stage gain is

$$\dot{K} = \dot{V}_{\text{out}}/\dot{V}_{\text{in}} = \dot{V}_{\text{L}}/\dot{V}_1 = (n/m)\,(\dot{V}_2/\dot{V}_1) \tag{3.9}$$

The ratio $\dot{V}_2/\dot{V}_1$ can be found from the second line of Eqs. (3.1) of the two-port, on substituting for $\dot{I}_2$ from Eq. (3.6):

$$-\dot{V}_2\dot{Y}'_{\Sigma} = \dot{Y}_{21}\dot{V}_1 + \dot{Y}_{22}\dot{V}_2$$

Hence,

$$\dot{V}_2/\dot{V}_1 = -\dot{Y}_{21}/(\dot{Y}_{22} + \dot{Y}'_{\Sigma}) \tag{3.10}$$

On substituting Eq. (3.10) in Eq. (3.9), we obtain

$$\dot{K} = -(n/m)\,\dot{Y}_{21}/(\dot{Y}_{22} + \dot{Y}'_{\Sigma}) \tag{3.11}$$

In view of Eq. (3.7), we may write Eq. (3.11) as

$$\dot{K} = -mn\dot{Y}_{21}/\dot{Y}_{\text{eq}} = -mn\dot{Y}_{21}R_{\text{eq}}/(1 + j\xi) \tag{3.12}$$

where

$$\begin{aligned}\dot{Y}_{\text{eq}} &= \dot{Y}_{\Sigma} + m^2\dot{Y}_{22} = \dot{Y}_{\text{ckt}} + m^2\dot{Y}_{22} + n^2\dot{Y}_{\text{L}} \\ &= G_{\text{eq}}\,(1 + j\xi)\end{aligned} \tag{3.13}$$

is the equivalent admittance of the tuned circuit. In Eq. (3.13),

$$G_{\text{eq}} = 1/R_{\text{eq}} = G_0 + m^2G_{22} + n^2G_{\text{L}} \tag{3.14}$$

is the equivalent tuned-circuit conductance at resonance, and

$$\xi = y/d_{\text{eq}} = (1/d_{\text{eq}})\,(\omega/\omega_0 - \omega_0/\omega)$$

is the generalized detuning (amount off resonance).

From comparison of Eqs. (3.12) and (2.10) it is seen that they differ solely in the sign and value of admittances: in Eq. (3.12), the admittance $1/|\,\dot{Z}_{\text{A}}\,|$ is replaced by the admittance $\dot{Y}_{21}$. Therefore, the relations derived in analysis of input circuits on the basis of Eq. (2.10) may be extended to include amplifiers.

As follows from Eq. (3.12), the magnitude of the gain is

$$K = mn\,|\,\dot{Y}_{21}\,|\,R_{\text{eq}}/(1 + \xi^2)^{1/2} \tag{3.15}$$

On setting $\xi = 0$, the gain at resonance is found to be

$$K_0 = mn\,|\,\dot{Y}_{21,0}\,|\,R_{\text{eq}} = mn\,|\,Y_{21,0}\,|/(G_0 + m^2G_{22} + n^2G_{\text{L}}) \tag{3.16}$$

Let us find the optimal values of m and n at which K_0 is a maximum at a specified total damping factor d_{eq}. If

$$D = d_{\text{eq}}/d_{\text{ckt}} = G_{\text{eq}}/G_0$$

then, subject to Eq. (3.14)

$$G_{\text{eq}} = DG_0 = m^2G_{22} + n^2G_{\text{L}} + G_0 \tag{3.17}$$

Therefore, in agreement with Eq. (3.16),

$$K_0 = mn\,|\,\dot{Y}_{21,0}\,|/DG_0 \tag{3.18}$$

On finding m from Eq. (3.17) and substituting it in Eq. (3.18), we obtain

$$K_0 = \frac{n\,|\dot{Y}_{21,0}|}{DG_0}\left[\frac{(D-1)\,G_0 - n^2G_{\mathrm{L}}}{G_{22}}\right]^{1/2} \tag{3.19}$$

The gain will be a maximum at

$$n_{\mathrm{opt}} = \left(\frac{D-1}{2}\,\frac{G_0}{G_{\mathrm{L}}}\right)^{1/2} \tag{3.20}$$

Subject to Eq. (3.20), it follows from Eq. (3.17) that

$$m_{\mathrm{opt}} = \left(\frac{D-1}{2}\,\frac{G_0}{G_{22}}\right)^{1/2} \tag{3.21}$$

On substituting Eqs. (3.20) and (3.21) in Eq. (3.18), we find

$$K_{0,\mathrm{max}} = \frac{|\dot{Y}_{21,0}|}{2\,(G_{\mathrm{L}}G_{22})^{1/2}}\,(1 - 1/D) \tag{3.22}$$

It is seen from Eqs. (3.20) and (3.21) that the gain is a maximum when the tuned circuit is shunted to an equal extent by both the amplifying element and the load, that is, when

$$m^2G_{22} = n^2G_L = (D-1)\,G_0/2$$

When the natural damping factor of the tuned circuit is low, that is, when $D \gg 1$, the gain achieves its limit:

$$K_{0,\mathrm{lim}} = |\dot{Y}_{21,0}|/2\,(G_LG_{22})^{1/2}$$

If the natural damping factor of the tuned circuit is close to the total damping factor chosen such that the specified selectivity is obtained, that is, when it is relatively large, the gain is low because with D tending to unity, K_0 tends to zero. It is seen therefore that one should seek to make the natural damping factor of the tuned circuit as small as possible.

We obtain from Eqs. (3.15) and (3.16)

$$K_0/K = 1/\gamma = \frac{|\dot{Y}_{21,0}|}{|\dot{Y}_{21}|}\,(1+\xi^2)^{1/2} \tag{3.23}$$

Not very far away from resonance it is legitimate to neglect changes in $|\dot{Y}_{21}|$. Then we have from Eq. (3.23)

$$1/\gamma = [1 + (2\Delta f/f_0d_{\mathrm{eq}})^2]^{1/2}$$

Hence the bandwidth at the specified flatness of the frequency response is found to be

$$B_\gamma = f_0d_{\mathrm{eq}}\,(1/\gamma^2 - 1)^{1/2} \tag{3.24}$$

With the frequency response flat to within $\gamma = 1/\sqrt{2} = 0.707$, or 3 dB, the bandwidth is found to be

$$B_{3\text{-dB}} = f_0 d_{\text{eq}}$$

In view of Eq. (3.4), the phase-vs-frequency response of the amplifier is defined as

$$-\varphi_{\text{amp}} = \text{arc tan}\ \xi + \text{arc tan}\ \omega\tau_{21} \tag{3.25}$$

Let us find the input admittance of the amplifier at terminals *1-1* in the circuit of Fig. 3.6. From the first line of Eqs. (3.1) we get

$$\dot{Y}_{\text{in}} = \dot{I}_1/\dot{V}_1 = \dot{Y}_{11} + \dot{Y}_{12}\dot{V}_2/\dot{V}_1 \tag{3.26}$$

On substituting for $\dot{V}_2/\dot{V}_1$ in Eq. (3.26) from Eq. (3.10), we get

$$\dot{Y}_{\text{in}} = \dot{Y}_{11} - \dot{Y}_{12}\dot{Y}_{21}/(\dot{Y}_{22} + \dot{V}'_{\Sigma}) \tag{3.27}$$

In view of Eqs. (3.11) and (3.12),

$$\dot{Y}_{\text{in}} = \dot{Y}_{11} = \dot{Y}_{12}m\dot{K}/n = \dot{Y}_{11} - m^2\dot{Y}_{12}\dot{Y}_{21}/\dot{Y}_{\text{eq}} \tag{3.28}$$

In Eqs. (3.26) through (3.28), the second term is due to the internal feedback admittance $\dot{Y}_{12}$.

3.4. The Effect of Internal Feedback on the Properties of a Tuned Amplifier

In amplifiers, feedback may take place over supply circuits, connecting leads, and the internal feedback admittance of the amplifying device. The first two forms of feedback may in principle be lowered to a tolerable value through the choice of a proper circuit configuration and component design.

In a common-source FET amplifier, the internal feedback is determined by the transfer capacitance, $C_{12} = C_{\text{GD}}$. In a BJT amplifier, $\dot{Y}_{12}$ is a complex quantity defined as

$$-\dot{Y}_{12} = G_{12} + j\omega C_{12} = |\dot{Y}_{12}| \exp(j\varphi_{12}) \tag{3.29}$$

where

$$|\dot{Y}_{12}| = [G_{12}^2 + (\omega C_{12})^2]^{1/2} \tag{3.30}$$

$$\varphi_{12} = \text{arc tan}\ (\omega C_{12}/G_{12}) = \text{arc tan}\ \omega\tau_{12} \tag{3.31}$$

Consider an amplifier in which the input tuned circuit has the configuration shown in Fig. 3.7 where, for ease of presentation, the internal feedback elements are shown as an external circuit. Admittance $\dot{Y}_{12}$ gives rise to a current, $\dot{I}$, at the amplifier input, which is equivalent to the existence of an admittance, $\dot{Y}_{\text{in},fb}$, called the input dynamic admittance. From Eq. (3.28), we get

$$\dot{Y}_{\text{in},fb} = \frac{-m_2^2\dot{Y}_{21}\dot{Y}_{21}}{\dot{Y}_{\text{eq}2}} = \frac{-m_2^2\dot{Y}_{12}\dot{Y}_{21}R_{\text{eq}}}{1+j\xi} \tag{3.32}$$

In view of Eqs. (3.2) and (3.29),

$$Y_{\text{in},fb} = m_2^2R_{\text{eq}2}\,|\,\dot{Y}_{12}\dot{Y}_{21}\,|\,\exp(j\varphi)/(1+j\xi) \tag{3.33}$$

where

$$\varphi = \varphi_{12} + \varphi_{21} = \text{arc tan}\,[\,\omega(\tau_{12} - \tau_{21})/(1 + \omega^2\tau_{12}\tau_{21})] \tag{3.34}$$

is the argument of the product $Y_{12}Y_{21}$. In Eq. (3.34), φ_{12} and φ_{21} are

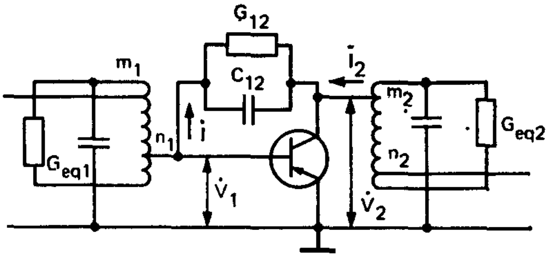


Fig. 3.7

defined by Eqs. (3.31) and (3.4). Using the Euler equation and rearranging Eq. (3.33), we get

$$\dot{Y}_{\text{in},fb} = m_2^2R_{\text{eq}2}\,|\,\dot{Y}_{12}\dot{Y}_{21}\,|\,\frac{\cos\varphi + \xi\sin\varphi}{1+\xi^2}$$
$$+\,jm_2^2R_{\text{eq}2}\,|\,\dot{Y}_{12}\dot{Y}_{21}\,|\,\frac{\sin\varphi - \xi\cos\varphi}{1+\xi^2} = G_{\text{in},fb} + jB_{\text{in},fb} \tag{3.35}$$

It is seen from Eq. (3.35) that $\dot{Y}_{\text{in},fb}$ is a complex admittance. It may be resolved into two conductances, $G_{\text{in},fb1}$ and $G_{\text{in},fb2}$, and two susceptances, $B_{\text{in},fb1}$ and $B_{\text{in},fb2}$, such that

$$\begin{aligned}
G_{\text{in},fb1} &= m_2^2R_{\text{eq}2}\,|\,\dot{Y}_{12},\ \dot{Y}_{21}\,|\,\cos\varphi/(1+\xi^2)\\
G_{\text{in},fb2} &= m_2^2R_{\text{eq}2}\,|\,\dot{Y}_{12}\dot{Y}_{21}\,|\,\sin\varphi\xi/(1+\xi^2)\\
B_{\text{in},fb1} &= \omega C_{\text{in},fb1} = m_2^2R_{\text{eq}2}\,|\,\dot{Y}_{12}\dot{Y}_{21}\,|\,\sin\varphi/(1+\xi^2)\\
B_{\text{in},fb2} &= \omega C_{\text{in},fb2} = -m_2^2\dot{R}_{\text{eq}2}\,|\,\dot{Y}_{12}\dot{Y}_{21}\,|\,\cos\varphi\xi/(1+\xi^2)
\end{aligned} \tag{3.36}$$

Plots of the functions $G_{\text{in},fb} = f(\xi)$ and $C_{\text{in},fb} = f(\xi)$ are shown in Fig. 3.8. These components of the input dynamic admittance shunt the input tuned circuit (see Fig. 3.7), thus affecting its frequency response.

In a FET amplifier, $G_{12} \approx 0$, $\tau_{12} \approx \infty$, $\tau_{21} \approx 0$, the transconductance g_m is real, and $\varphi = \pi/2$. Therefore, $G_{\text{in,fb}} \approx 0$, $B_{\text{in,fb2}} \approx 0$, and

$$G_{\text{in,fb2}} = \omega C_{12} g_m m_2^2 R_{\text{eq2}} \xi/(1 + \xi^2)$$
$$B_{\text{in,fb1}} = \omega C_{\text{in,fb1}} = \omega C_{12} g_m m^2 R_{\text{eq2}}/(1 + \xi^2)$$

To begin with, consider the effect of these components. Suppose that the input tuned circuit is tuned to the same frequency as the output tuned circuit. If all the components of the input dynamic admittance were independent of frequency, the frequency response of the tuned circuit would have the shape shown by the solid line in Fig. 3.9. Actually, $G_{\text{in.fb2}}$ and $B_{\text{in,fb1}}$ vary with frequency. At frequencies below resonance, the conductance is negative and causes the gain to increase (the dashed curve in Fig. 3.9). This can be explained as follows. Below resonance, the impedance of the output tuned circuit is inductive in its effect. For this reason, $\dot{V}_2$ (see Fig. 3.7) leads $\dot{I}_2$ by an angle close to 90° (Fig. 3.10). The voltage $\dot{V}_2$ gives rise to the current $\dot{I}$ through the capacitance C_{12}, leading the voltage by another 90°. Since $\dot{I}_2$ is in phase with $\dot{V}_1$, the phase shift between

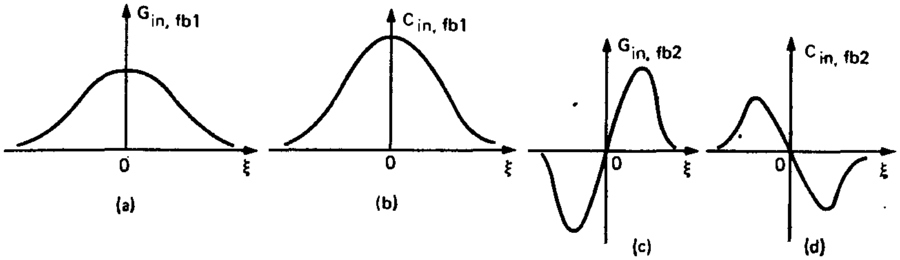


Fig. 3.8

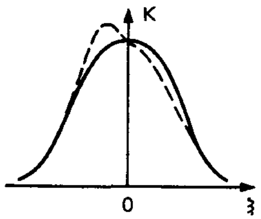


Fig. 3.9

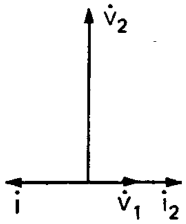


Fig. 3.10

$\dot{V}_1$ and $\dot{I}$ is 180°, which is equivalent to a negative conductance. It makes up for the losses in the input tuned circuit by increasing the voltage. That is, positive feedback is produced in the amplifier.

Above resonance, $G_{\mathrm{in.fb2}}$ is positive. It couples losses into the tuned circuit, thereby bringing down the gain. This is another way of saying that negative feedback takes place.

The effect of $G_{\mathrm{in,fb1}}$ on the frequency response of the input tuned circuit consists in that a decrease in frequency causes the total tuned-circuit capacitance to fall and the resonance frequency to rise.

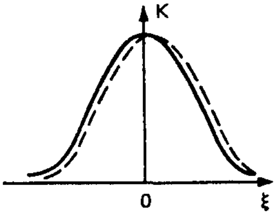


Fig. 3.11

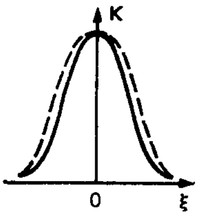


Fig. 3.12

Actually, the amount off resonance is greater than the decrease in frequency, therefore the gain rolls off more rapidly (the dashed curve to the left of the axis of ordinates in Fig. 3.11). As the frequency rises, the total capacitance decreases, and the resonant frequency goes up. The tuned circuit 'servos' to the desired frequency, as it were, the actual amount off resonance decreases, and the gain turns out to be greater than it would be in the absence of feedback (the dashed curve to the right of the axis of ordinates in Fig. 3.11).

In a BJT amplifier, C_{12} causes similar changes in the frequency response. Since, however, the forward and reverse admittances Y_{12} and $\dot{Y}_{21}$ are complex, all the four components of the input dynamic admittance are present (see Fig. 3.8).

At resonance, $C_{\mathrm{in,fb2}}$ is zero. As the frequency goes down, $C_{\mathrm{in,fb2}}$ increases, the resonant frequency decreases, and the actual amount off resonance also decreases. As the frequency goes up, $G_{\mathrm{in,fb2}}$ turns negative, the total capacitance decreases, the resonant frequency rises, and this leads to a decreased amount off resonance and an increased gain (the dashed curve in Fig. 3.12). Far away from resonance, $C_{\mathrm{in,fb2}}$ decreases and does not affect the frequency response any longer.

The conductance $G_{\mathrm{in,fb1}}$ decreases on moving away either side from resonance. As this happens, the Q-factor of the input tuned circuit is improved, the gain rises to the right and left of the resonance frequency, and the resonance curve widens at the top (the dashed curve in Fig. 3.12).

To sum up, feedback distorts the resonance curve. In fact, the amplifier may even jump into oscillations due to the negative conductance $G_{\text{in,fb2}}$.

3.5. The Condition for the Stability of an Amplifier

An amplifier will go oscillating, if the following equalities are satisfied:

$$\left.\begin{aligned} B_{\text{eq1}} + n_1^2 B_{\text{in,fb}} = 0 \\ G_{\text{eq1}} + n_1^2 G_{\text{in,fb}} = 0 \end{aligned}\right\} \qquad (3.37)$$

The first line defines the balance of phases, and the second line, the balance of amplitudes. An amplifier will not jump into oscillations if its input tuned-circuit conductance, considering the feedback, is positive

$$G_{\text{eq1}} + n_1^2 G_{\text{in,fb}} > 0$$

However, the absence of oscillations does not mean that the amplifier shows a consistent performance.

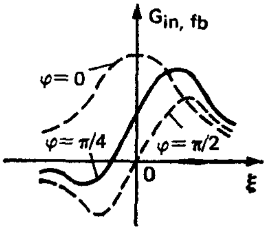


Fig. 3.13

Let us introduce the stability factor

$$k_{\text{st}} = (G_{\text{eq1}} + n_1^2 G_{\text{in,fb}})/G_{\text{eq1}} \qquad (3.38)$$

If $k_{\text{st}} = 0$, the amplifier may go oscillating. At $k_{\text{st}} = 1$, there is no feedback, and the amplifier has a maximum stability. Ordinarily, it is taken that $k_{\text{st}} = 0.8$ or 0.9. Then variations in the gain and bandwidth due to feedback will not exceed 10 to 20%. The stability of an amplifier improves as k_{st} moves closer to unity.

A similar reasoning applies in the case of negative feedback: the performance of the amplifier ought not to change materially. Therefore, it is usual to take $k_{\text{st}} = 1.1$ to 1.2.

Let us find the condition for the stability of an amplifier with a specified stability margin. It follows from Eq. (3.35) that

$$G_{\text{in,fb}} = m_2^2 R_{\text{eq2}} \mid \dot{Y}_{12} \dot{Y}_{21} \mid g(\varphi, \xi) \qquad (3.39)$$

where

$$g(\varphi, \xi) = (\cos\varphi + \xi \sin\varphi)/(1 + \xi^2) \qquad (3.40)$$

is a function defining the dependence of $G_{\text{in,fb}}$ on the detuning (amount off resonance) ξ and on the phase angle φ. A plot of this function is shown in Fig. 3.13, where the solid line illustrates the dependence of $G_{\text{in,fb}}$ on ξ, and the dashed curves, its dependence on the phase. As the phase angle varies, it approaches $G_{\text{in,fb1}}$ or $G_{\text{in,fb2}}$ (see Fig. 3.8*a* or *c*). With positive feedback, $G_{\text{in,fb}}$ is negative, whereas with negative feedback it is positive. On substituting for $G_{\text{in,fb}}$

from Eq. (3.39) in Eq. (3.38), we obtain

$$k_{st} = 1 + n_1^2 m_2^2 R_{eq2} R_{eq1} \mid \dot{Y}_{12} \dot{Y}_{21} \mid g\,(\varphi,\ \xi) \tag{3.41}$$

It follows from Eq. (3.41) that

$$m_2^2 R_{eq2} \mid Y_{21} \mid \; = \; \mid k_{st} - 1 \mid / n_1^2 R_{eq2} \mid \dot{Y}_{21} \mid \mid g\,(\varphi,\ \xi) \mid \tag{3.42}$$

In Eq. (3.42), the absolute values $\mid k_{st} - 1 \mid$ and $\mid g\,(\varphi,\ \xi) \mid$ are used in order to combine the two cases of feedback in one equation, because with positive feedback $k_{st} < 1$ and $g\,(\varphi,\ \xi) < 0$, whereas with negative feedback $k_{st} > 1$ and $g\,(\varphi,\ \xi) > 0$. On multiplying both sides of Eq. (3.42) by $n_2^2 \mid \dot{Y}_{21} \mid R_{eq2}$ and solving for $K_0 = m_2 n_2 R_{eq2} \mid \dot{Y}_{21} \mid$ we get

$$K_{0,st} \leqslant (n_2/n_1) \left(\frac{\mid k_{st} - 1 \mid}{\mid g\,(\varphi, \xi) \mid} \, \frac{\mid \dot{Y}_{21} \mid}{\mid \dot{Y}_{12} \mid} \, \frac{R_{eq2}}{R_{eq1}} \right)^{1/2} \tag{3.43}$$

An amplifier will be stable if $K_0 \leqslant K_{0,st}$. Given identical tuned circuits (such that $R_{eq1} = R_{eq2}$) and $n_1 = n_2$, Eq. (3.43) takes the form

$$K_{0.st} \leqslant \left(\frac{\mid k_{st} - 1 \mid}{\mid g\,(\varphi,\ \xi) \mid} \, \frac{\mid \dot{Y}_{21} \mid}{\mid \dot{Y}_{12} \mid} \right)^{1/2} \tag{3.44}$$

For a greater and more stable amplification, the amplifying device should be chosen such that the ratio $\mid \dot{Y}_{21} \mid / \mid \dot{Y}_{12} \mid$ is a maximum. This ratio, $A_{amp} = \mid \dot{Y}_{21} \mid / \mid \dot{Y}_{12} \mid$, characterizes both the capabilities of the amplifying device and the parasitic feedback. Amplification will be feasible, if $A_{amp} > 1$.

Analysis of Eq. (3.40) shows that the minimum and maximum of $g\,(\varphi,\ \xi)$ do exist:

$$g\,(\varphi,\ \xi)_{min} = -\,(1 - \cos\varphi)/2 \tag{3.45}$$

$$g\,(\varphi,\ \xi)_{max} = (1 + \cos\varphi)/2 \tag{3.46}$$

In a BJT amplifier, the amount of negative feedback exceeds that of positive feedback. In a FET amplifier, the two forms of feedback are equal ($\varphi = \pi/2$). In analysing the effect of feedback, one should proceed from the highest absolute value of $g\,(\varphi,\ \xi)$. In agreement with Eqs. (3.46) and (3.44)

$$K_{0,st} = [2 \mid k_{st} - 1 \mid A_{amp}/(1 + \cos\varphi)]^{1/2} \tag{3.47}$$

Consider several special cases.

1. $\cos\varphi = 1$ and $\varphi = 0$, which; according to Eq. (3.34), corresponds to the equality $\tau_{12} = \tau_{21}$. Then it follows from Eq. (3.47) that

$$K_{0,st} = (\mid k_{st} - 1 \mid A_{amp})^{1/2} \tag{3.48}$$

When $\tau_{12} = \tau_{21}$, only negative feedback exists. The amplifier cannot jump into oscillations. Its performance will not vary markedly, and its gain will not exceed the value defined by Eq. (3.48). For transistors, it is usual that $\tau_{12} > \tau_{21}$.

2. $\cos \varphi = 0$ and $\varphi = \pi/2$. Then, as follows from Eq. (3.47),

$$K_{0,\mathrm{st}} = [(2 \mid k_{\mathrm{st}} - 1 \mid A_{\mathrm{amp}}]^{1/2} \tag{3.49}$$

This case corresponds to the inequalities

$$\left.\begin{array}{l} \omega^2 \ll \omega_{gm}^2 \\ (\omega C_{12})^2 \gg G_{12}^2 \end{array}\right\} \tag{3.50}$$

and

which are satisfied if the transistor is operating within the frequency range below the frequency at which its transconductance has a limiting value. Then Eq. (3.49) takes the form

$$K_{0,\mathrm{st}} = (\mid k - 1 \mid 2g_{\mathrm{m}}/\omega C_{12})^{1/2} \tag{3.51}$$

At $k_{\mathrm{st}} = 0.9$ or $k_{\mathrm{st}} = 1.1$,

$$K_{0,\mathrm{st}} = 0.45\,(g_{\mathrm{m}}/\omega C_{12})^{1/2} \tag{3.52}$$

For a tube amplifier, the stable gain is given by Eqs. (3.51) and (3.52) in which it is taken that $C_{12} = C_{\mathrm{pg}}$.

In a multistage amplifier, the output tuned circuit of a preceding stage is the input one for the next stage and is shunted by its input admittance. Because of this, the equivalent admittance of the preceding stage is changed, and this entails a more pronounced change in the input admittance and the parameters of its input tuned circuit. For this reason, multistage amplifiers are less stable than single-stage amplifiers. If each stage satisfies the condition defined by Eq. (3.52), the stability will be impaired only slightly.

In designing an amplifier it should be borne in mind, however, that in a multistage high-gain amplifier external feedback is quite likely to appear due to the inductances and capacitances directly between the wiring and structural components in the input and output stages.

3.6. Ways and Means of Improving the Stability of Tuned Amplifiers

The stability of a tuned amplifier can be improved in several ways which may be classed into passive and active. Passive techniques boil down to reducing the actual gain so as to satisfy the inequality

$$K_0 \leqslant K_{0,\mathrm{st}} \tag{3.53}$$

To achieve this, it may suffice to reduce the tapping-down factor or the tuned-circuit impedance R_{eq}.

Let us find the tapping-down factor m_2 (see Fig. 3.7) so as to satisfy the condition stated in Eq. (3.53). For this purpose, we use Eqs. (3.16) and (3.43):

$$m_2 n_2 \mid Y_{21,0} \mid \dot{R}_{eq2} \leqslant (n_2/n_1)\left[\frac{|k_{st}-1|}{g(\varphi,\xi)}\,\frac{|\dot{Y}_{21}|}{|\dot{Y}_{12}|}\,\frac{R_{eq2}}{R_{eq1}}\right]^{1/2}$$

Hence,

$$m_2 \leqslant (1/n_1 R_{eq})\left(\frac{|k-1|}{|g(\varphi,\xi)|\;|\dot{Y}_{12}\dot{Y}_{21}|}\right)^{1/2} \qquad (3.54)$$

where $R_{eq} = (R_{eq1}R_{eq2})^{1/2}$, and n_1 is known from the calculation of the preceding stage or input circuit. In view of Eq. (3.46) and assuming that the condition stated in Eq. (50) is satisfied, Eq. (3.54) takes the form

$$m_2 \leqslant (1/n_1 R_{eq})\left(\frac{2\,|k_{st}-1|}{|\dot{Y}_{12}\dot{Y}_{21}|}\right)^{1/2} \qquad (3.55)$$

The tapping-down factor n_2 is taken such that one obtains the required damping factor

$$d_{eq} = d_{ckt} + m_2^2 \rho G_{22} + n_2^2 \rho G_L$$

Hence,

$$n_2 = \left[\frac{(d_{eq} - d_{ckt}) - m_2^2 \rho G_{22}}{\rho G_L}\right]^{1/2} \approx [(d_{eq} - n_{ckt})/\rho G_L]^{1/2}$$

Active techniques of stability improvement offer a means to raise $K_{0,st}$ and, thus, to realize the potential capabilities of the amplifying device. These techniques include internal feedback neutralization by an opposite external feedback and the cascade connection of the active elements.

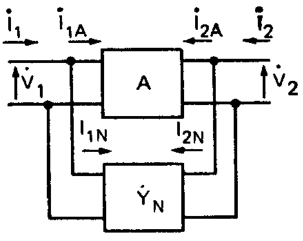


Fig. 3.14

The internal feedback of an amplifying device can be neutralized by incorporating suitable networks. This removes the constraint imposed on the gain by the inequality (3.53) and offers an opportunity to obtain the maximum gain defined by Eq. (3.22). Several neutralizing circuit configurations are known. A shunt-neutralized amplifier is a parallel connection of two four-terminal networks: the amplifying device A and a neutralizing circuit of admittance $\dot{Y}_N$, as shown in Fig. 3.14. Let us find the resultant parameter Y_{12} for the two parallel-connected two-ports. By definition,

$$\dot{Y}_{12L} = \frac{\dot{I}_1}{\dot{V}_2}\bigg|_{\dot{V}_1=0} = \frac{\dot{I}_{1amp} + \dot{I}_{1N}}{\dot{V}_2} = \dot{Y}_{12} + \dot{Y}_N$$

There is no feedback if

$$\dot{Y}_{12L} = \dot{Y}_{12} + \dot{Y}_N = 0$$

Hence, the condition for neutralization takes the form

$$\dot{Y}_N = -\dot{Y}_{12}$$

In consequence, the neutralizing circuit should be analogous in its properties to the $\dot{Y}_{12}$ circuit. The feedback voltage should be applied over the neutralizing circuit to the amplifier input in anti-phase with that which reaches the input via the internal feedback path. This purpose can be served by an autotransformer or transformer phase inverter.

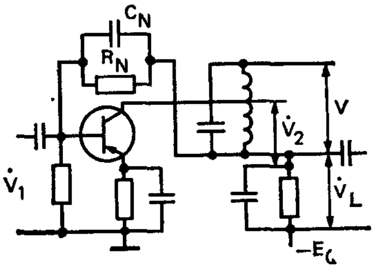


Fig. 3.15

Figure 3.15 shows an amplifier with an autotransformer phase inverter and a shunt neutralizing circuit $R_N C_N$. This circuit can provide exact neutralization within a frequency range where G_{12} and C_{12} are practically constant. In the case of BJTs, G_{12}, and C_1, are frequency-dependent, which is why no neutralization is used in broadband and bandpass amplifiers.

As an alternative, series neutralization can be used. It is effected by a circuit in which C_N and R_N are connected in series. It effects exact neutralization at the frequency at which the series neutralizing circuit is equivalent to a shunt neutralizing circuit. This can be achieved within the passband of a selective amplifier tuned to fixed frequencies. Series neutralization is convenient to use in cases where there should be no resistive coupling between the output and input circuits of the amplifier. This is possible because C_N also doubles as a d.c. blocking capacitor.

As has been noted, amplifier stability can be improved by a cascade connection of two amplifying devices. In this arrangement, the output of the first amplifying device is connected to the input of the second directly, omitting any frequency-dependent networks. This serves to minimize the effect of internal feedback because the feedback admittance is determined by the inverse admittance of the two amplifying devices.

In the 1940s it was widely practiced to arrange amplifiers so that a neutralized common-cathode input stage was followed by a common-grid output stage. This arrangement came to be known as the *cascode amplifier*. Since the advent of transistors, the term 'cascode' has been used to describe any amplifiers which have no frequency-dependent coupling between cascade-connected transistors. For purposes of analysis, it is convenient to treat such a connection as a single stage in which the two amplifying devices have been replaced by an equivalent two-port as shown in Fig. 3.16, having equivalent parameters defined as

$$\begin{aligned}\dot{Y}_{11\mathrm{eq}} &= \dot{Y}'_{11} - \dot{Y}'_{21}\dot{Y}'_{12}/(\dot{Y}'_{22} + \dot{Y}''_{11})\\ \dot{Y}_{22\mathrm{eq}} &= \dot{Y}''_{22} - \dot{Y}''_{21}\dot{Y}''_{12}/(\dot{Y}'_{22} + \dot{Y}''_{11})\cdot\\ \dot{Y}_{21\mathrm{eq}} &= -\dot{Y}'_{21}\dot{Y}''_{21}/(\dot{Y}'_{22} + \dot{Y}''_{11})\\ \dot{Y}_{12\mathrm{eq}} &= -\dot{Y}'_{12}\dot{Y}''_{12}/(\dot{Y}'_{22} + \dot{Y}''_{11})\end{aligned} \tag{3.56}$$

Subject to the above relations, the design equations (3.15), (3.16), (3.19), and (3.22) remain valid.

In BJT amplifiers, it is usual to have a neutralized common-emitter input stage followed by a common-emitter output stage (the CE-CE configuration) or a neutralized common-emitter input stage followed by a common-base output stage (the CE-CB configuration). The CE-CE configuration is used at frequency $f_0 \leqslant 1$ or 2 MHz, such as in the i.f. amplifiers of broadcast receivers. The CE-CB configuration is used at higher frequencies, notably in the HF and VHF bands, and in broadband equipment.

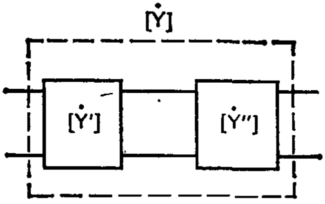


Fig. 3.16

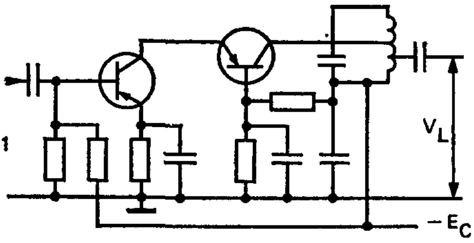


Fig. 3.17

In FET amplifiers, good performance is shown by the common source-common gate (CS-CG) configuration, although the common source-common drain (CS-CD) configuration is also used.

Cascode amplifiers provide a high stable gain without a need to use neutralization. The parameter

$$A_{\mathrm{amp,eq}} = |\dot{Y}_{21\mathrm{eq}}|/|\dot{Y}_{12\mathrm{eq}}|$$

of a cascode amplifier can be found by Eq. (3.56). Notably, for a CE-CE amplifier using identical transistors,

$$A_{\mathrm{amp,eq}} = |\dot{Y}_{21}|^2/|\dot{Y}_{12}|^2$$

whereas for the CE-CB configuration it is

$$A_{\mathrm{amp,eq}} = |\dot{Y}_{21}|^2/|\dot{Y}_{12}(\dot{Y}_{12} + \dot{Y}_{22})|$$

which is substantially greater than in the case of a common-emitter amplifier. The stable gain is defined by Eq. (3.47) considering the new value of $A_{\mathrm{amp,eq}}$. Figure 3.17 shows the circuit of a series-fed CE-CB cascode amplifier. Wide use is made of integrated-circuit (IC) cascode amplifiers.

3.7. Performance of a Tuned Amplifier over a Frequency Range

Consider how the gain at resonance varies with frequency in the amplifiers shown in Figs. 3.1 through 3.3.

In FET amplifiers, the loss coupled into the tuned circuit because the amplifying devices shunt the tuned circuit does not usually exceed the value tolerated from the viewpoint of the required selectivity. Tapping-down on the tuned-circuit coil is only necessary in order to satisfy the stability condition, Eq. (3.53). The tapping-down factor m in a tapped-coil-coupled amplifier should be chosen in agreement with Eq. (3.55). For an amplifier tuned by a variable capacitor, such as shown in Fig. 3.1, the gain-vs-frequency relation at $n = 1$ has, in agreement with Eq. (3.16), the following form:

$$K_0 = m\,|\dot{Y}_{21,0}|\,R_{\mathrm{eq}} = m g_{\mathrm{m}} \omega_0 L_{\mathrm{ckt}} Q_{\mathrm{eq}} \tag{3.57}$$

Since, practically, $|\dot{Y}_{21,0}| = g_{\mathrm{m}} =$ const and $Q_{\mathrm{eq}} =$ const, it follows that K_0 increases with increasing frequency.

In BJT amplifiers, use is often made of double-tapped-coil coupling, as shown in Fig. 3.2. The dependence of gain on frequency is now described by a relation far more elaborate than Eq. (3.57). Subject to Eq. (3.3) it follows from Eq. (3.16) that

$$K_0 = \frac{m n g_m R_{\mathrm{eq}}}{[1 + (\omega_0/\omega_{\mathrm{gm}})^2]^{1/2}} = \frac{m n g_m \omega_0 L_{\mathrm{ckt}} Q_{\mathrm{eq}}}{[1 + (\omega_0/\omega_{gm})^2]^{1/2}} \tag{3.58}$$

where

$$Q_{\mathrm{eq}} = 1/d_{\mathrm{eq}} = \frac{1}{[d_{\mathrm{ckt}} + m^2 \omega_0 L_{\mathrm{ckt}} G_{22} + n^2 \omega_0 L_{\mathrm{ckt}} G_{\mathrm{L}}} \tag{3.59}$$

In such a case, m and n are independent of frequency:

$$m = (L_1 + M_1)/L_{\mathrm{ckt}}$$
$$n = (L_2 + M_2)/L_{\mathrm{ckt}}$$

where M_1 = mutual inductance between L_1 and the remaining turns of the tuned-circuit coil L_{ckt}

M_2 = mutual inductance between L_2 and the remaining turns of L_{ckt}

Both $\dot{Y}_{21,0}$ and Q_{eq} are frequency-dependent. If $\omega_{gm}^2 \gg \omega_0^2$, the transconductance is practically constant. As the frequency increases, the Q-factor of the tuned circuit decreases owing to the loss coupled in from the stage output ($m^2\rho G_{22}$) and from the load ($n^2\rho G_L$). If the tuned circuit is loosely coupled to the amplifying devices, the Q-factor decreases at not a very high rate whereas K_0 rises but not so fast as in Eq. (3.57). The bandwidth increases with increasing frequency:

$$B_{0.7} = f_0 d_{eq} = f_0 \left(d_{ckt} + m^2\omega_0 L_{ckt} G_{22} + n^2\omega_0 L_{ckt} G_L\right)$$

The circuit shown in Fig. 3.3 is an amplifier with a transformer-coupled tuned circuit. It is widely used in transistor receivers. Its equivalent circuit appears in Fig. 3.18*a*. Let the amplifying element be represented by an equivalent current generator $\dot{Y}_{21}\dot{V}_1$ with an output conductance G_{22} and an output capacitance C_{out} which includes, in addition to C_{22}, also the wiring capacitance of the output circuit and the capacitance of the coupling coil, L_c. The resonance frequency of the coupling circuit is

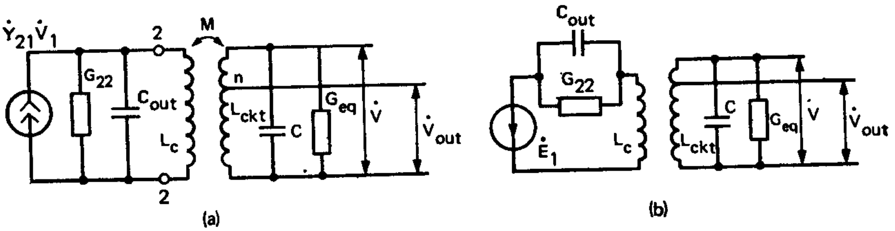


Fig. 3.18

$$\omega_c = 1/(L_c C_{out})^{1/2} \tag{3.60}$$

According to Thévenin's theorem, the circuit in Fig. 3.18*a* can be re-arranged to have the configuration shown in Fig. 3.18*b* where the emf $\dot{E}_1$ is found as the open-circuit voltage between terminals *2-2* in Fig. 3.18*a*:

$$\dot{E}_1 = \dot{Y}_{21}\dot{V}_1/(G_{22} + j\omega C_{out}) \approx \dot{Y}_{21}\dot{V}_1/j\omega C_{out} \tag{3.61}$$

Here and elsewhere, G_{22} is neglected because $G_{22} \ll \omega C_{out}$. In view of Eqs. (3.61) and (3.60), the current traversing the coupling coil L_c is

$$\begin{aligned} \dot{I}_c &= \dot{E}_1/(j\omega L_c + 1/j\omega C_{out}) \\ &= \dot{Y}_{21}\dot{V}_1/[1 - (\omega/\omega_c)^2] \end{aligned} \tag{3.62}$$

In the case at hand, the tuned-circuit capacitance is

$$C = C_{\text{ckt}} + n^2 C_{\text{in}} + C_{\text{M1}}$$

The emf induced in the tuned circuit is

$$\dot{E}_2 = j\omega M \dot{I}_{\text{c}}$$

Considering Eq. (3.62), the voltage across the tuned circuit $L_{\text{ckt}}C$ at resonance is

$$V_0 = |\, E_{2,0} \,| Q_{\text{eq}} = |\, \dot{I}_{\text{c}} \,| \omega_0 M Q_{\text{eq}} = \frac{|\, \dot{Y}_{21,0} | V_1 M Q_{\text{eq}} \omega_0}{|1 - (\omega_0/\omega_{\text{c}})^2|}$$

Hence,

$$K_0 = V_{\text{ont}}/V_1 = nV_0/V_1 = \frac{n \,|\, \dot{Y}_{21,0} \,|\, (M/L_{\text{ckt}})\, R_{\text{eq}}}{1 - (\omega_0/\omega_{\text{c}})^2} \tag{3.63}$$

Equation (3.63) will be the same as Eq. (3.16), if we denote

$$m\,(\omega_0) = \frac{M/L_{\text{ckt}}}{|\, 1 - (\omega_0/\omega_{\text{c}})^2 \,|} \tag{3.64}$$

The amplifier will operate in any one of several modes depending on the value of the ratio $\omega_0/\omega_{\text{c}}$. If $\omega_{\text{c}}^2 \gg \omega_0^2$ then, as is seen from Eq. (3.64), $m \approx M/L_{\text{ckt}}$. The gain at resonance depends on frequency in the same manner as with double tapped-coil coupling:

$$K_0 = n\,(M/L_{\text{ckt}}) \,|\, \dot{Y}_{21,0} \,|\, R_{\text{eq}} \tag{3.65}$$

which means that K_0 increases with increasing frequency.

When $\omega_{\text{c}}^2 \ll \omega_0^2$,

$$K_0 \approx n\,(M/L_{\text{ckt}}) \,|\, \dot{Y}_{21,0} \,|\, R_{\text{eq}} \omega_{\text{c}}^2/\omega_0^2 \tag{3.66}$$

Hence, for $|\, \dot{Y}_{21,0} \,| = g_{\text{m}} = \text{const}$ and $Q_{\text{eq}} = \text{const}$,

$$K_0 \approx nMg_{\text{m}}Q_{\text{eq}}\omega_{\text{c}}^2/\omega_0 = \text{const}/\omega_0 \tag{3.67}$$

If one takes into account variations in $|\, \dot{Y}_{21,0} \,|$ with frequency, then Eq. (3.66) may be re-cast as

$$K_0 = \frac{nMg_{\text{m}}Q_{\text{eq}}\omega_{\text{c}}^2}{\omega_0\,[1 + (\omega_0/\omega_{g\text{m}})^2]^{1/2}} \tag{3.68}$$

where Q_{eq} is defined by Eq. (3.59).

In BJT amplifiers, it is of interest to consider internal capacitive coupling between the tuned circuit and the input of the next stage, similar to that used in the input circuit shown in Fig. 2.16. It is advantageous in that it preserves the high selectivity of the tuned circuit at the high-frequency end of the frequency range owing to a decrease in n and, as a consequence, a reduction in the shunting effect of the input admittance with increasing frequency. Figure 3.19 shows a modified circuit configuration of an amplifier which uses a combination of internal capacitive coupling (C_c) and transformer coupling (L_c) from the tuned circuit to the input of the next stage. With this form of coupling, the Q-factor can be maintained or even increased somewhat with an increase in frequency. When $\omega_c > \omega_0$, K_0 varies only slightly over the band.

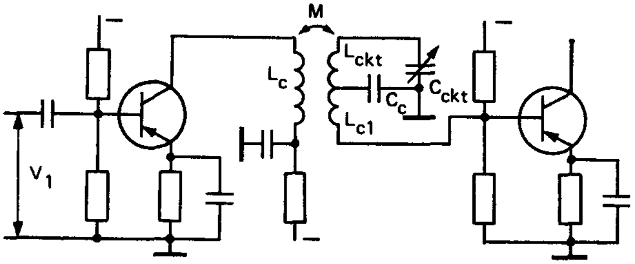


Fig. 3.19

3.8. The Combined Noise Factor of a Tuned Amplifier Together with Its Input Circuit

In simplified form, the typical arrangement of a part of a receiver is shown in Fig. 3.20. A signal source of complex admittance $Y_s =$ $G_s + jB_s$ and the input of an amplifying stage are connected to a tuned circuit such that the respective tapping-down factors are $m = V_s/V$ and $n = V_1/V$.

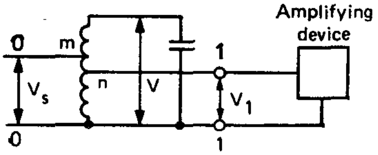


Fig. 3.20

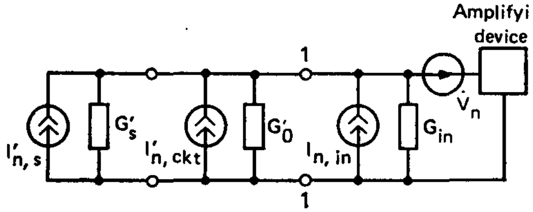


Fig. 3.21

Let us find the combined noise factor of the input circuit and of the first stage of the receiver. For this purpose we draw an equivalent circuit, as shown in Fig. 3.21, following the recommendations given in Sec. 1.7. Since the circuits are tuned to the signal frequency, Fig. 3.21 shows solely the conductance. Thermal noise originating in the signal source and the tuned circuit is represented by equivalent current generators transferred to the input terminals of the amplifying device (terminals *1-1*):

$$I'_{\mathrm{n,s}} = (4kTBG'_{\mathrm{s}})^{1/2} \qquad (3.69)$$

$$I'_{\mathrm{n,ckt}} = (4kTBG'_{0})^{1/2} \qquad (3.70)$$

Here, $G'_{\mathrm{s}} = (m^2/n^2)G_{\mathrm{s}}$ and $G'_0 = G_0/n^2$ are the conductances of the signal source and tuned circuit, transferred to terminals *1-1*.

The noise originating in the amplifying device is represented by an equivalent noise current generator, $I_{\mathrm{n,in}}$, Eq. (1.11), and an equivalent noise voltage generator, V_{n}, Eq. (1.10). The current associated with the noise voltage V_{n} is

$$I_{\mathrm{n}} = V_{\mathrm{n}}G_{\Sigma} = (4KTBR_{\mathrm{n}})^{1/2}\,(G'_{\mathrm{s}} + G'_0 + G_{\mathrm{in}}) \qquad (3.71)$$

where

$$G_{\Sigma} = G'_{\mathrm{s}} + G'_0 + G_{\mathrm{in}} = (m^2G_{\mathrm{s}} + G_0 + n^2G_{\mathrm{in}})/n^2 \qquad (3.72)$$

By definition, the noise factor is

$$N = \frac{\sum\limits_i I^2_{\mathrm{n}i}}{I'^2_{\mathrm{n,s}}} = 1 + \frac{I'^2_{\mathrm{n,ckt}} + I^2_{\mathrm{n,in}} + I^2_{\mathrm{n}}}{I'^2_{\mathrm{n,s}}} \qquad (3.73)$$

Substitution of Eqs. (3.69) through (3.71) and (1.11) in Eq. (3.73) gives

$$N = 1 + G'_0/G'_{\mathrm{s}} + t_{\mathrm{in}}G_{\mathrm{in}}/G'_{\mathrm{s}} + (R_{\mathrm{n}}/G'_{\mathrm{s}})\,(G'_{\mathrm{s}} + G'_0 + G_{\mathrm{in}})^2 \qquad (3.74)$$

Let us find the optimal value of G'_{s} for which the noise factor is a minimum. To this end, we solve the equation

$$dN/dG'_{\mathrm{s}} = 0$$

for G'_{s}, and get

$$G_{\mathrm{s,opt}} = (G'_0 + G_{\mathrm{in}})\left[1 + \frac{G'_0 + t_{\mathrm{in}}G_{\mathrm{in}}}{R_{\mathrm{n}}\,(G'_0 + G_{\mathrm{in}})^2}\right]^{1/2} \qquad (3.75)$$

On substituting Eq. (3.75) in Eq. (3.74), we obtain the minimal noise factor

$$N_{\mathrm{min}} = 1 + 2R_{\mathrm{s}}\,(G'_0 + G_{\mathrm{in}} + G'_{\mathrm{s,opt}}) \qquad (3.76)$$

Noting that

$$G'_{\mathrm{s,opt}} = m^2_{\mathrm{opt}}G_{\mathrm{s}}/n^2$$

we can use Eq. (3.75) to find the tapping-down factor for which the noise factor is a minimum:

$$m_{\mathrm{opt,n}} = m_{\mathrm{m}}\left[1 + \frac{G_0/n + t_{\mathrm{in}}G_{\mathrm{in}}}{R_{\mathrm{n}}\,(G_0/n + G_{\mathrm{in}})^2}\right]^{1/4} \qquad (3.77)$$

where

$$n_m = \left[\frac{G'_0 + G_{\mathrm{in}}}{G_{\mathrm{s}}/n^2}\right]^{1/2} = [(G_0 + n^2G_{\mathrm{in}})/G_{\mathrm{s}}]^{1/2} \qquad (3.78)$$

is the tapping-down factor providing for a proper match between the signal source and the first stage of the receiver.

A plot relating the noise factor, as defined by Eq. (3.76), to m appears in Fig. 3.22. For ease of comparison, this figure also shows a plot of the input tuned circuit gain at resonance, K_0, as a function of m. It is seen that $m_{\mathrm{opt,n}} > m_{\mathrm{m}}$. This difference exists when the amplifying device has a low noise level and the total noise is mainly due to the signal source and the input circuit. The point is that the impedance coupled by the signal source into the tuned circuit rapidly increases with increasing m, so that the thermal noise originating in the tuned circuit proper decreases in comparison with the thermal noise originating in the signal source. With a high noise level in the amplifying device, the noise factor will be a minimum at perfect match, $m_{\mathrm{opt,n}} \approx m_{\mathrm{m}}$.

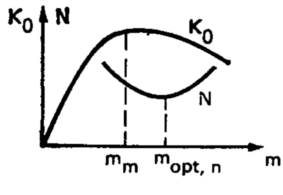


Fig. 3.22

Mismatch at the receiver input may be undesirable in operation with a tuned antenna, as a feeder echo might appear. In operation with a tuned antenna it is customary to match the antenna to the feeder and the feeder to the receiver input, thus ensuring a travelling-wave condition. The condition for match is given by the equality

$$G'_{\mathrm{s}} = G'_0 + G_{\mathrm{in}} \tag{3.79}$$

Then the tapping-down factors m and n are as given by Eqs. (2.31) and (2.32). The noise factor under conditions of match can be found on substituting (3.79) into (3.74):

$$N_{\mathrm{m}} = 2 + \frac{G_{\mathrm{in}}}{G'_0 + G_{\mathrm{in}}}(t_{\mathrm{in}} - 1) + 4R_{\mathrm{n}}(G'_0 + G_{\mathrm{in}}) \tag{3.80}$$

For FET amplifiers ($t_{\mathrm{in}} = 1$, $n = 1$, and $G_{\mathrm{in}} \ll G_0$), it follows from Eq. (3.80) that

$$N_{\mathrm{m}} = 2 + 4R_{\mathrm{n}}G_0$$

The condition of match at receiver input is of primary importance, the more so that optimum mismatch seldom gives a tangible reduction in the noise factor. In the general case, in order to bring down the noise factor, the amplifying device should be chosen such that the product $R_{\mathrm{n}}G_{\mathrm{in}}$ has the lowest possible value. That is why FETs are a better choice for the early stages of a receiver than BJTs.

3.9. Low-Noise Microwave Amplifiers

In the LF, MF, HF and VHF bands, one does not strive to obtain a noise factor lower than is naturally provided by state-of-the-art transistors, because external noise and interference would stand in the way of receiving weak signals all the same. In the UHF, SHF bands and at the low-frequency end of the EHF band, weak signals are amplified by special low-noise amplifiers — the *maser amplifier* and the *parametric amplifier*. Ordinarily, they are arranged as *regenerative* or (though more seldom) as *travelling-wave amplifiers*. In maser amplifiers, the signal field is amplified by drawing upon the molecular energy of matter. Such amplifiers have a very low noise level, but they are rather elaborate in arrangement. At this writing, they are mainly used in deep-space communication systems and in radio astronomy.

Parametric amplifiers have a somewhat higher noise level. For their operation they depend on the conversion of the energy supplied by what is called a 'pumping' or 'pump' oscillator into the energy of the signal being amplified. The conversion is effected by nonlinear reactive elements, most often varactors.

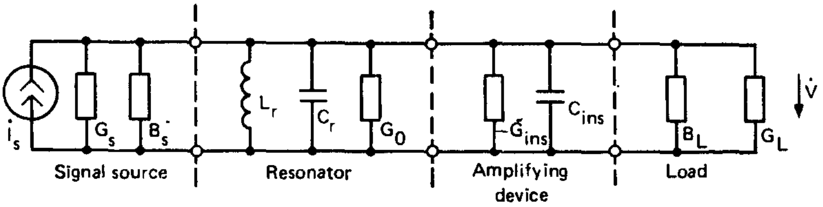


Fig. 3.23

Also falling in the category of regenerative amplifiers are those using tunnel diodes which display negative resistance owing to their behaviour within the tunnel effect region.

A general equivalent circuit of a regenerative amplifier is shown in Fig. 3.23. The amplifier is seen to include a resonator, $L_r C_r$, of equivalent loss conductance G_0, to which are connected a signal source and a load, both transferred to the resonator. The effect of the energy source ensuring amplification is depicted in the form of a negative conductance, $-G_{ins}$, coupled into the resonator, and insertion capacitance, C_{ins}. At resonance, the susceptance, that is, the reactive component of admittance, of the resonator is zero.

The power gain can be found as the ratio of the load power to the available signal-source power

$$K_P = P_L/P_{s,av} \tag{3.81}$$

The load power is given by

$$P_L = V^2 G_L = (I_L/G_\Sigma)^2 G_L = \frac{I_s^2 G_L}{(G_s + G_0 + G_L - G_{ins})^2} \tag{3.82}$$

and the available signal-source power is given by Eq. (1.14). On substituting Eqs. (1.14) and (3.82) in Eq. (3.81), we get

$$K_P = \frac{4 G_s G_L}{(G_s + G_0 + G_L - G_{ins})^2} = \frac{4 G_s G_L}{G_{eq}^2 (1 - q)^2} \tag{3.83}$$

where

$$G_{eq} = G_s + G_0 + G_L$$

and $q = G_{ins}/G_{eq}$ is the regeneration factor. With q tending to unity, the power gain tends to infinity, but practically the power gain is never greater than 10 to 20 dB because the amplifier goes oscillating.

The bandwidth of a regenerative amplifier between 3-dB points is

$$B_{3\text{-dB}} = f_0 d_r = f_0 \rho G_\Sigma = f_0 d_{eq} (1 - q) \qquad (3.84)$$

where $d_{eq} = \rho G_{eq}$ and $f_0 d_{eq}$ are the damping factor and bandwidth of the resonator (tuned circuit) without regeneration. As is seen from Eqs. (3.83) and (3.84), an increase in gain is accompanied by a reduction in bandwidth.

A major advantage of the above type of amplifier is the low level of internal noise. This is mainly thermal noise which can be brought down by cooling. In order to obtain a low noise factor, it is essential to prevent load noise from finding its way into the amplifier because this form of noise would be amplified along with the received signal and there would be no improvement in the real receiver sensitivity. Transfer of load noise to the amplifier resonator can be prevented by isolators and circulators. Isolators are used in feed-through amplifiers. In them, the signal picked up by the antenna passes through the first isolator to enter a resonator, and the amplified signal is routed to the load via a second isolator which serves to prevent load noise from finding its way back into the resonator.

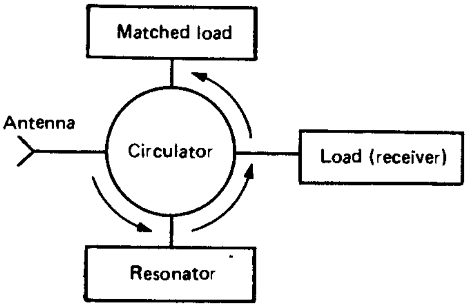


Fig. 3.24

In reflection-type amplifiers, such as shown in block form in Fig. 3.24, the signal picked up by the antenna is fed to the resonator via a circulator. Amplified in the resonator, the signal goes via the circulator to the input of the next stage. Inside the circulator, power can only follow the path indicated by the arrows. Load noise is absorbed by a matched resistive network and cannot reach the resonator. The figure shows a four-arm circulator. Wide use is also made of three-arm or Y-circulators.

In the 1970s, low-noise BJT amplifiers were developed for operation at frequencies to 30 GHz and higher, with a performance comparable with that of tunnel-diode amplifiers. Among their advantages are high reliability, low cost, simple design, unidirectional amplification, a relatively low noise factor, a negligible warm-up time, simplicity of servicing, and ease of miniaturization. That is why tunnel-diode amplifiers are being used on a limited scale. In more detail, parametric amplifiers are discussed in Chap. 4.

3.10. Intermediate-Frequency Amplifiers with Single-Tuned Stages Tuned to the Same Frequency

Intermediate-frequency (i.f.) amplifiers are ordinarily tuned to a fixed frequency and may contain several stages so as to achieve a specified gain.

Consider an i.f. amplifier containing N identical stages. To this end, we will use the results obtained in Sec. 3.3. For an N-stage amplifier, the gain is

$$K_N(\omega) = [K(\omega)]^N = \left[\frac{mn\,|\dot{Y}_{21}|\,R_{\text{eq}}}{(1+\xi^2)^{1/2}}\right]^N$$

and the gain at resonance is

$$K_N(\omega_0) = [K(\omega_0)]^N = (mn\,|\dot{Y}_{21,0}|\,R_{\text{eq}})^N \tag{3.85}$$

Therefore,

$$\frac{K_N(\omega_0)}{K_N(\omega)} = 1/\gamma_N \left[\frac{|\dot{Y}_{21,0}|}{|\dot{Y}_{21}|}(1+\xi^2)\right]^N \tag{3.86}$$

If $|\dot{Y}_{21,0}| \approx |\dot{Y}_{21}|$, Eq. (3.86) takes the form

$$1/\gamma_N = [(1+\xi^2)^{1/2}]^N \tag{3.87}$$

If the frequency response of the amplifier is flat to witbin γ_N, its bandwidth will be

$$B_{\gamma_N} = f_0 d_{\text{eq}}\,[(1/\gamma_N^2)^{1/N} - 1]^{1/2}$$

If the frequency response of the amplifier is flat to within $\gamma_N = 1/\sqrt{2} = 0.707 = 3$ dB, its bandwidth between the 3-dB points will be

$$B_{0.7N} = f_0 d_{\text{eq}}\,(2^{1/N} - 1)^{1/2} = B_{0.7}/\psi_1(N) \tag{3.88}$$

where

$$B_{0.7} = f_0 d_{\text{eq}}$$

is the 3-dB bandwidth of each stage, and

$$\psi_1(N) = 1/(2^{1/N} - 1)^{1/2}$$

is a function of the number of stages.

It is seen from Eq. (3.88) that in order to obtain the specified bandwidth for the entire amplifier, one has to extend the bandwidth of each stage. To this end, the damping factor of each tuned circuit is taken as

$$d_{\text{eq}} = B_{0.7N}\psi_1(N)/f_0 \tag{3.89}$$

The bandwidth ratio (or the bandwidth shape factor) of the amplifier

$$K_{\mathrm{br},\gamma} = B_{\gamma N}/B_{0.7N} = [(1/\gamma_N^2)^{1/N} - 1]^{1/2}/(2^{1/N} - 1)^{1/2} \quad (3.90)$$

varies with the number of stages. For a single-stage amplifier, $K_{\mathrm{br},0.1} \approx 10$. As the number of stages is increased, the bandwidth ratio improves, but there is a limit for improvement. For example, with N tending to infinity, $K_{\mathrm{br},0.1} \approx 2.6$.

The phase response of a multistage amplifier is

$$\varphi_{\mathrm{amp},N} = N\varphi_{\mathrm{amp}}$$

A high gain is easy to obtain with narrowband amplifiers, because it is solely limited by the condition for stability. To demonstrate, if the value of m has been chosen so as to obtain a stable gain in agreement with Eqs. (3.54) and (3.55), and the value of n bas been chosen so as to obtain a specified damping factor, a reduction in amplifier bandwidth will be accompanied by a reduction in its gain. Under the conditions we have been examining, the gain is independent within certain limits from the tuned-circuit capacitance. To demonstrate, as the tuned-circuit capacitance is increased to some critical value, C_{cr}, the resonance impedance R_{eq} also decreases. Therefore, one has simultaneously to increase both m and n so that the gain will remain unchanged so long as $m < 1$. That is why in narrowband amplifiers one may increase the tuned-circuit capacitance without detriment to gain, which fact has a wholesome effect on stability.

In broadband amplifiers, it is usual that $m = 1$. Then Eq. (3.85) may be re-cast as

$$\begin{aligned} K_N(\omega_0) &= (n\,|\dot{Y}_{21,0}|\,R_{\mathrm{eq}})^N = (n\,|\dot{Y}_{21,0}|\rho d_{\mathrm{eq}})^N \\ &= (n\,|\dot{Y}_{21,0}|\,/2\pi\,CB_{0.7})^N \quad (3.91) \end{aligned}$$

where

$$\begin{aligned} C &= C_{\mathrm{ckt}} + C_{\mathrm{out}} + n^2C_{\mathrm{in}} + C_{\mathrm{w}} \\ &= C_{\Sigma} + n^2C_{\mathrm{in}} \quad (3.92) \end{aligned}$$

It is seen from Eq. (3.91) that the gain decreases with increasing tuned-circuit capacitance and bandwidth. This explains why it is difficult to obtain a high gain in broadband amplifiers. The capacitance may be decreased as far as

$$C = C_{\mathrm{out}} + n^2C_{\mathrm{in}} + C_{\mathrm{w}}$$

Its further reduction is limited by the likely impairment in amplifier performance stability.

It is seen from Eqs. (3.91) and (3.92) that the gain depends on the pping-down factor n in two ways. The optimal value of n is given by

$$n_{\text{opt}} = (C_\Sigma / C_{\text{in}})^{1/2}$$

Any further increase in amplifier bandwidth can be obtained by shunting the tuned circuit with a resistor, R_{sh}.

Let us re-arrange Eq. (3.91) subject to Eq. (3.88):

$$K_N(\omega_0) = \left[\frac{n \mid \dot{Y}_{21,0} \mid}{2\pi C B_{0.7N} \psi_1(N)}\right]^N = K_{\text{eq}}^N / \psi_N(N) \qquad (3.93)$$

where $K_{\text{eq}} = n \mid \dot{Y}_{21,0} \mid / 2\pi C B_{0.7N}$ is the gain of one stage with a bandwidth specified for the multistage amplifier, and $\psi_N(N) = [\psi_1(N)]^N$. The factor $\psi_1(N)$ shows that an increase in the number of stages, with the bandwidth held constant, causes the gain of each stage to decrease. This is because, in order to maintain the specified bandwidth, one has to increase the damping factor of each tuned circuit in proportion to $\psi_1(N)$, as defined by Eq. (3.89). As the stage number N is increased, the gain $K_N(\omega_0)$ at first increases until N exceeds some critical value, after which $K_N(\omega_0)$ begins to decrease. That is why in broadband amplifiers with identically tuned resonant circuits a high gain at a specified bandwidth is not always achievable. A high gain-bandwidth product can be obtained in stagger-tuned amplifiers or amplifiers containing bandpass filters.

3.11. The Amplifier with a Double-Tuned Filter

Amplifiers with a double-tuned filter come in several arrangements. In the most commonly used schemes, the tuned circuits are inductively and externally capacitively coupled. The tuned circuits are coupled to the amplifying devices by a tapped coil or by a capacitive divider.

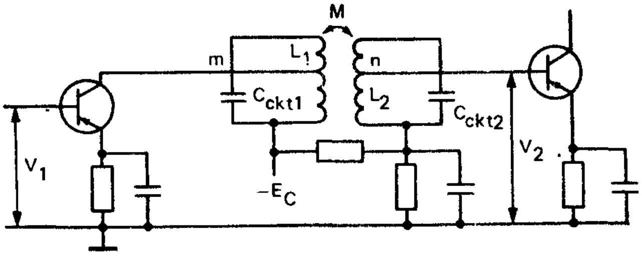


Fig. 3.25

Consider the arrangement in which the tuned circuits are inductively coupled as shown in Fig. 3.25. The basic results will hold for other arrangements as well. Consider an equivalent circuit in which the output of the amplifying device is replaced by an equivalent current generator $\dot{Y}_{21}\dot{V}_1$ of conductance G_{out} and capacitance C_{out}, and the input of the next stage is replaced by conductance G_{in} and capacitance C_{in}. The equivalent circuit is shown in Fig. 3.26 where

$$C_1 = C_{ckt1} + m^2C_{out} + C_{w1}$$
$$C_2 = C_{ckt2} + n^2C_{in} + C_{w2}$$

are the overall capacitances, and

$$G_{eq1} = G_{01} + m^2G_{out}$$
$$G_{eq2} = G_{02} + n^2G_{in}$$

are the overall conductances.

Proceeding from Norton's and Thévenin's theorems let us replace the equivalent current generator $m\dot{Y}_{21}\dot{V}_1$ by an equivalent voltage generator $\dot{E}_1$, as shown in Fig. 3.27. Here, the emf $\dot{E}_1$ is found as the open-circuit voltage between terminals *1-1*:

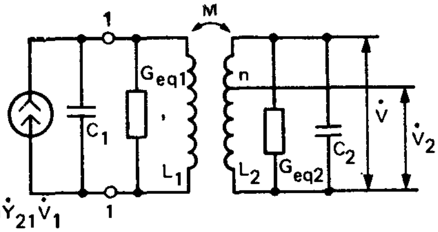


Fig. 3.26

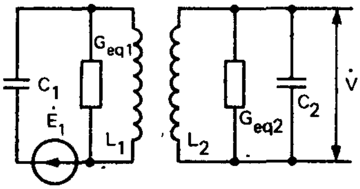


Fig. 3.27

$$\dot{E}_1 = m\dot{Y}_{21}\dot{V}_1/j\omega C_1$$

Knowing the filter transmission gain

$$\dot{K}_f = \dot{V}/\dot{E}_1$$

it is an easy matter to determine the amplifier gain

$$\begin{aligned}\dot{K} = \dot{V}_2/\dot{V}_1 = n\dot{V}/\dot{V}_1 &= (mn\dot{Y}_{21}/j\omega C_1)\dot{K}_f \\ &= -j\,(\omega_0/\omega)\,mn\dot{Y}_{21}\rho_1\dot{K}_f \qquad (3.94)\end{aligned}$$

where

$$\rho_1 = 1/\omega_0C_1$$

is the characteristic impedance of the first tuned circuit.

Equation (3.94) holds for an amplifier with a filter containing any number of tuned circuits (with a proportionate K_f).

The phase response of an amplifier with a double-tuned filter is determined by those of the filter and the amplifying device. In contrast to a single-tuned amplifier, there is an additional phase shift of $-\pi/2$. The magnitude of the gain is

$$K = (\omega_0/\omega)mn \mid \dot{Y}_{21} \mid \rho_1 K_f$$

Close to resonance (such that $\omega_0/\omega \approx 1$), the frequency response of the amplifier is basically determined by that of the filter

$$K = mn \mid \dot{Y}_{21} \mid \rho_1 K_f \tag{3.95}$$

Expressions for K_f are supplied by the theory of linear circuits. Given identical tuned circuits, K_f for a double-tuned amplifier is given by

$$K_f = \beta/d_{eq}\,[(1 + \xi^2 - \beta^2)^2 + 4\beta^2]^{1/2} \tag{3.96}$$

where

$$\beta = k_c/d_{eq}$$

In view of Eq. (3.96), we may re-write Eq. (3.95) as

$$K \approx mn \mid \dot{Y}_{21} \mid R_{eq}\beta/[(1 + \xi^2 - \beta^2)^2 + 4\beta^2]^{1/2}$$

The gain of an N-stage amplifier is given by

$$K_N(\omega) = \{mn \mid \dot{Y}_{21} \mid R_{eq}\beta/[(1 + \xi^2 - \beta^2)^2 + 4\beta^2]^{1/2}\}^N \tag{3.97}$$

At resonance ($\xi = 0$),

$$K_N(\omega) = [mn \mid \dot{Y}_{21,0} \mid R_{eq}\beta/(1 + \beta^2)]^N \tag{3.98}$$

If the amplifying device has been chosen with an ample frequency margin, it follows from Eqs. (3.97) and (3.98) that the frequency response of the amplifier is

$$\frac{K_N(\omega_0)}{K_N(\omega)} = \left\{\frac{[(1+\xi^2-\beta^2)^2+4\beta^2]^{1/2}}{1+\beta^2}\right\}^N$$

The shape of the frequency response depends on β. With β less than unity, it has one hump. With β equal to unity (critical coupling), the frequency response curve has a flat top. With β greater than unity, it has two humps.

The frequency response curve comes closest to a square one when the valley between the humps lies within the allowable variations in gain over the bandwidth. From the view-point of alignment, it is most convenient to use filters in which the tuned circuits are critically coupled ($\beta = 1$). Then, the phase response will also be close to a linear one.

3.12. Amplifiers with a Multisection Filter

Multisection filters (MSFs) serve to give high selectivity and, at the same time, a flat gain within the specified bandwidth. It is warranted to use them if the i.f. amplifier is built around an IC amplifier module of a sufficiently high gain so that there is no need for separate amplifying stages.

MSFs may be of the *LC* type widely varying in complexity, electromechanical types, and piezoceramic types. An MSF is mainly responsible for the frequency response of the entire i.f. section. Where a need exists for additional stages, their bandwidth is made greater than that of the MSF so as not to impair the overall frequency response.

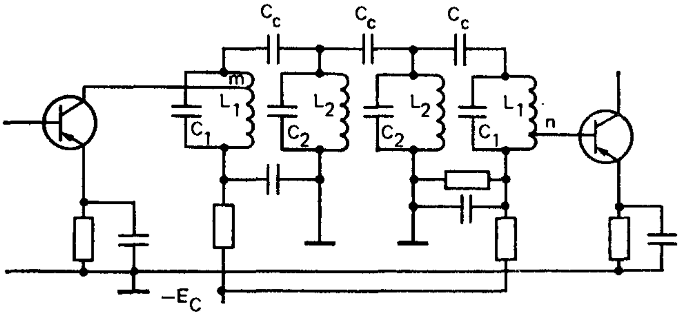


Fig. 3.28

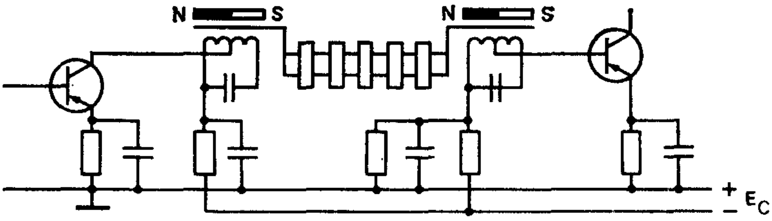


Fig. 3.29

Now that selectivity is concentrated in one stage, the frequency response of the i.f. section has a greater stability against variations in temperature and supply voltages. Where selectivity is provided by a chain of stages, the frequency response is less stable because of the spread in parameters between the individual transistors.

An example of a multisection *LC* filter is shown in Fig. 3.28, whereas Fig. 3.29 shows an electromechanical filter. It has an input magnetostriction transducer which converts electrical oscillations into mechanical vibrations, a mechanical filter, and an output transducer which converts mechanical vibrations into electrical oscillations. Magnetostriction is that property of certain ferromagnetic materials (such as nickel and Permalloy) which causes them to shrink or expand when placed in a magnetic field. The filter consists of a number of mechanical resonators in the form of plates, rods or discs linked by elastic couplers. Mechanical vibrations of the input transducer induce vibrations in the mechanical resonators each of which resonates similarly to a high-Q resonant circuit. The final resonator induces vibrations in the output resonator which converts these vibrations into electrical oscillations by virtue of the inverse magnetostrictive effect. Such filters have a nearly square frequency response curve, small size, and good temperature stability.

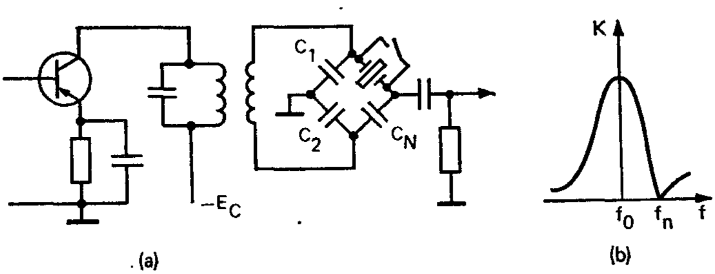


Fig. 3.30

Where it is necessary to obtain a very narrow bandwidth (of the order of several hundred or tens of hertz), resort is made to quartz-crystal filters, such as shown in Fig. 3.30*a*. The filtering action of a quartz-crystal resonator is based on a high rate of rolloff in its impedance (and, in consequence, a high rate of cutoff) within a narrow band of frequencies on either side of the resonance frequency. As a way of neutralizing the capacitance of the crystal holder, the filter is arranged in a bridge circuit. The bridge arms are formed by capacitors C_1, C_2 and C_N, and by the capacitance of the crystal holder. At frequency f_n (Fig. 3.30*b*) where the impedance of the quartz crystal is capacitive in its effect, the bridge is at balance, and the output voltage is practically zero. On moving away from that frequency, the balance is upset, and an output voltage is developed, being a maximum at the frequency of series resonance, f_0, of the quartz plate. Quite often, a quartz-crystal filter may contain several quartz-crystal resonators.

The piezoelectric effect occurs not only in single crystals, but also in polycrystalline substances. Among the latter are piezoceramic materials which can be fabricated into resonators of any required shape and size, suitable for the manufacture of miniature resonators. They are low in cost and small in size and can be arranged into elaborate filters having a frequency response with a very abrupt cutoff. As an example, Fig. 3.31 shows a *ladder-type filter*. As compared with quartz, piezoceramic materials have a lower temperature and time stability and suffer higher losses. On the other hand, they can provide a relative (percentage) bandwidth of the order of 0.1%.

In the HF and VHF bands, use is made of *acoustic surface wave filters*. Such a filter consists of a piezoelectric substrate (quartz, lithium niobate, lithium tantalate, or bismuth germanate) onto which film transducers in the shape of interdigitated combs are deposited by photolithography, as shown in Fig. 3.32. When a signal is applied to the input transducer, an acoustic wave is induced in the interdigitated structure by virtue of the piezoelectric effect. This wave propagates in two directions away from the input transducer. In one direction it dies out in an absorbing medium, whereas in the other direction it reaches the output transducer where it undergoes the reverse piezoelectric effect. Acoustic surface wave filters fall in the class of *transversal filters*.

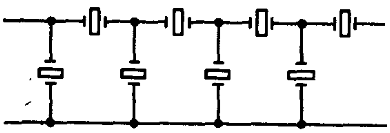


Fig. 3.31

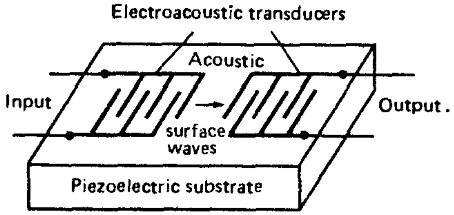


Fig. 3.32

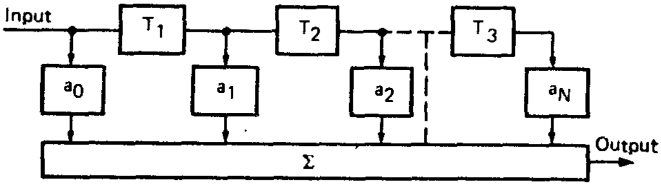


Fig. 3.33

Signal filtering may be treated as the addition of delayed signals which appropriate weight coefficients. The signals are combined in phase within the pass band and in anti-phase within the stop (suppression or rejection) band of the filter. The arrangement of a transversal filter is shown in Fig. 3.33. The filter contains a delay line with N taps, each tap having a distinct weight coefficient, a_n. The sum of the weighted signals picked off the taps make up the output voltage. The interdigitated transducer electrodes deposited on a substrate may be looked upon as the taps of a delay line, and the buses as summators. In contrast to the classic transversal filter, an acoustic surface wave filter has two sets of taps from the delay line. Its response is determined by two (input and output) transducers which can be adjusted at will so as to obtain the desired response.

Acoustic surface wave filters are other than the minimum-phase type because signals travel more than one input-to-output path in them. In minimum-phase filters, the amplitude-vs-frequency and phase-vs-frequency (or, simply, phase) responses are related to each other in a unique manner. Therefore, one has to add a corrector so as to obtain a linear phase response, and this complicates the filter. In nonminimum-phase filters, the amplitude and phase responses are independent of each other. Therefore, one can obtain a nearly square amplitude response and a linear phase response within the band.

Acoustic surface wave filters are mostly used at frequencies from 30 to 800 MHz with a percentage bandwidth of 0.1 to 30%. They may also be used at frequencies from 1 MHz to 3 GHz, with the lower edge frequency of the range being limited by the substrate size and the upper edge frequency by the feasibility of resonator manufacture. Among the advantages offered by acoustic surface wave filters are high selectivity, small size, applicability of IC technology, and compatibility with IC modules. In large-scale production, one obtains high performance reproducibility and stability, a relatively low cost, and high reliability.

3.13. Alternative Designs of Bandpass Amplifiers

By virtue of the analogy that exists between mechanical and electrical resonant systems piezoelectric or mechanical filters made up of a number of coupled resonators possess the same properties

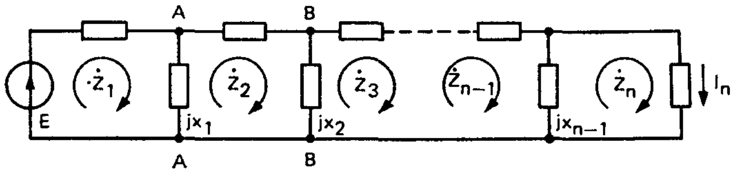


Fig. 3.34

as filters using electric resonators. In consequence, they may all be designed on the basis of a common theory. These filters may be represented by an equivalent electric circuit such as shown in Fig. 3.34 where jx_i are the reactances that couple the individual resonators.

Let us find the current in the nth resonator. This is an easy matter to do by invoking Thévenin's theorem. By disconnecting in turn the circuits to the right of points A-A, B-B, and so on, we find the emfs of the equivalent voltage generators as the open-circuit voltage between those points. The impedance of the ith equivalent generator (to the left of points A-A, B-B, etc.) is

$$\frac{jx_i(\dot{z}_i - jx_i)}{jx_i + (z_i - jx_i)} = jx_i + x_i^2/\dot{z}_i$$

It is easy to see that the second term on the right-hand side of the above equation is the impedance coupled into the $(i + 1)$st tuned circuit by the ith circuit.

By consecutively applying the above procedure, the current in the last tuned circuit is found to be

$$\dot{I}_n = E\frac{1}{\dot{z}_1}\, jx_1 \frac{1}{\dot{z}_2 + \frac{x_1^2}{\dot{z}_1}}\, jx_2 \frac{1}{\dot{z}_3 + \frac{x_2^2}{\dot{z}_2 + x_1^2/\dot{z}_1}}$$

$$\times\, jx_3 \ldots jx_{n-1} \frac{1}{\dot{z}_n + \frac{x_{n-1}^2}{\dot{z}_{n-1} + \frac{x_{n-1}^2}{\dot{z}_{n-1} + x_{n-2}^2/(\dot{z}_{n-2} + \ \ldots)}}} \qquad (3.99)$$

Knowing the current $\dot{I}_n$ in the last tuned circuit, it is an easy matter to find the output voltage of the filter.

When the fractions in the denominator of Eq. (3.99) are consecutively multiplied together from end to start, they cancel out, and the expression reduces to the following form:

$$\dot{I}_{\mathrm{n}} = (jx_1 jx_2 \ldots jx_{n-1})/(z_1 z_2 \ldots z_n + \ldots)$$

The above general expression can be used to derive the equations defining both the amplitude and phase characteristics of concrete filter configurations.

As a way of simplifying filter design and manufacture, it is common practice to tune all of their resonators to the same frequency, f_0. Allowing for the likely small difference in tuning, let us write

$$\dot{z}_i = \rho_i\,[d_i + j\,(y + \eta)]$$

where

$$\rho_i = 1/\omega_i C_i = \omega_i L_i$$

d = damping factor of the ith tuned circuit

$$y = 2\,(f - f_0)/f$$

On substituting for $\dot{z}_i$ and replacing jy by ζ, we get

$$\dot{I}_n = \varkappa/(\zeta^n + p_1^2\zeta^{n-1} + \ldots) \qquad (3.100)$$

The numerator is proportional to the product of the coupling coefficients of the individual resonators.

On equating the denominator of Eq. (3.100) to zero, we find the roots which are complex in the general case:

$$\zeta_i = -\delta_i + j\theta_i$$

They correspond to the roots of the characteristic equation in operational analysis of transients. The '—' sign in front of the real part of the root is indicative of the fact that self-sustained oscillations cannot be excited and built up in a passive network. θ_i may take on both positive and negative values. As a result, Eq. (3.100) may be re-cast as

$$\dot{I}_n = \dot{\varkappa}\frac{1}{\delta_1 + j(y-\theta_1)}\,\frac{1}{\delta_2 + j(y-\theta_2)}\,\cdots\,\frac{1}{\delta_n + j(y-\theta_n)} \quad (3.101)$$

It is seen from Eq. (3.101) that the amplitude and phase responses of the circuit in Fig. 3.34 are similar to the respective responses of a network which is a chain of single tuned circuits which, in the general case, somewhat differ in the resonance frequency (the difference being θ_i) and in the damping factors (represented by δ_i) of uncoupled tuned circuits (for example, isolated from one another by transistors or any other nonreciprocal elements). It follows then that it is feasible to synthesize a bandpass amplifier having the same characteristics with any combinations of coupled and single tuned circuits or their equivalents. For example, with $n = 2$, the same responses can be obtained by using in the amplifier a filter made up of either two coupled resonators or two uncoupled and stagger-tuned circuits (one each in the amplifying stages). With $n = 3$, the same responses can be obtained, using a triple-tuned filter or three uncoupled stagger-tuned circuits or (recalling the case with $n = 2$) with a single-tuned circuit in one stage and a double-tuned filter in the other.

The magnitude of the current defined by Eq. (3.101) is given by

$$I_n = \frac{\varkappa}{[(y^n + a_1 y^{n-1} + \ldots)^2 + (b_1 y^{n-1} + b_2 y^{n-2} \ldots)^2]^{1/2}}$$

or, in a different way,

$$I_n = \frac{\varkappa}{(y^{2n} + m y^{2n-2} + \ldots + q)^{1/2}}$$

The extrema of the amplitude-vs-frequency characteristic can be located on equating the derivative of the radicand in the denominator to zero and solving the resultant equation of degree $2n—1$. Accordingly, the number of roots may be equal to or less than $2n — 1$. Since with f tending to zero and with f tending to infinity, that is, with y^2 tending to infinity, the current $\dot{I}_n$ falls to zero, it is obvious that the outmost extrema of the amplitude response must be maxima, unless they are points of inflection. In consequence, given certain relationships between the circuit parameters, the frequency response may have n peaks and $n - 1$ valleys within the passband.

3.14. Performance Stability of I.F. Amplifiers

The gain, bandwidth, frequency response and phase response of an i.f. amplifier may all vary owing to the effect of destabilizing factors. Changes in ambient temperature and in supply voltages and currents for the amplifying devices are bound to cause changes in the input and output admittance, inverse admittance, and transconductance, thus leading to changes in amplifier performance.

The tuning of the tuned circuits and, as a consequence, the amplifier performance are most of all affected by changes in the input and output capacitances of the transistors because they are part of the respective tuned circuits. The overall tuned-circuit capacitance is given by

$$C = C_{\text{ckt}} + m^2 (C_{22} + \Delta C_{22}) + n^2 (C_{11} + \Delta C_{11})$$

in the case of single-tuned-circuit i.f. amplifiers, and

$$\begin{aligned} C_1 &= C_{\text{ckt1}} + m^2 (C_{22} + \Delta C_{22}) \\ C_2 &= C_{\text{ckt2}} + n^2 (C_{11} + \Delta C_{11}) \end{aligned}$$

in the case of double-tuned-circuit i.f. amplifiers.

Hence, changes in the capacitances are

$$\begin{aligned} \Delta C &= m^2 \Delta C_{22} + n^2 \Delta C_{11} \\ \Delta C_1 &= m^2 \Delta C_{22} \\ \Delta C_2 &= n^2 \Delta C_{11} \end{aligned}$$

The values of m and n are found from the specified damping factors of the tuned circuits and from the condition for stable amplification, as has been done in Sec. 3.6. In a double-tuned i.f. amplifier, the input and output capacitances are part of different tuned circuits, for which reason its stability is better than that of a single-tuned i.f. amplifier. The amplifier performance is stable if

$$\Delta C/C \leqslant \nu B/f_0$$

where ν is the limit of instability factor. Hence the overall capacitance of the i.f. amplifier tuned circuit should satisfy the condition

$$C \geqslant (\Delta C/\nu)\,(f_0/B)$$

The choice of tuned-circuit capacitances should satisfy the conditions defined below.

For single-tuned-circuit amplifiers,

$$C_{\mathrm{ckt}} \geqslant (\Delta C/\mathrm{v})\,(f_0/B) - m^2C_{22} - n^2C_{11} - C_{\mathrm{w}}$$

For double-tuned-circuit amplifiers,

$$C_{\mathrm{ckt}1} \geqslant (\Delta C_1/\mathrm{v})\,(f_0/B) - m^2C_{22} - C_{\mathrm{w}1}$$

$$C_{\mathrm{ckt}2} \geqslant (\Delta C_2/\mathrm{v})\,(f_0/B) - n^2C_{11} - C_{\mathrm{w}2}$$

In amplifiers using multisection filters (see Fig. 3.34), the capacitances of the amplifying devices solely affect the tuning of the first and last sections. Therefore, such amplifiers show a more stable performance than amplifiers in which the same tuned circuits or sections are used in different stages. If there is a gain margin, it will be a good plan to increase the capacitances because this will improve the performance stability.

Changes in the input and output admittances lead above all to changes in the bandwidth. The bandwidth is deemed stable if the following conditions are satisfied:

— for single-tuned-circuit i.f. amplifiers

$$\Delta G_{\mathrm{eq}}/G_{\mathrm{eq}} \leqslant \Delta B/B$$

— for double-tuned-circuit i.f. amplifiers

$$\Delta G_{\mathrm{eq}1}/G_{\mathrm{eq}1} \leqslant \Delta B/B$$

$$\Delta G_{\mathrm{eq}2}/G_{\mathrm{eq}2} \leqslant \Delta B/B$$

Here,

$$\Delta G_{\mathrm{eq}} = m^2\Delta G_{22} + n^2\Delta G_{11}$$

$$\Delta G_{\mathrm{eq}1} = m^2\Delta G_{22}$$

$$\Delta G_{\mathrm{eq}2} = n^2\Delta G_{11}$$

and $\Delta B/B$ is the maximum allowable fractional change in the bandwidth.

Variations in the transconductance lead to changes in gain. This can be avoided by temperature stabilization of supply voltages and a.c. negative feedback. These measures also serve to minimize the effect of changes in G_{11} and G_{12}.

The effect and control of internal feedback have been examined in Secs. 3.4 through 3.6.

3.15. Integrated-Circuit Amplifiers

In the past, radio equipment was assembled from discrete components. Since the 1970s it has largely been designed on the basis of complete functional subassemblies in the form of integrated circuits (ICs). As compared with discrete-component circuits, IC modules are far more reliable, draw less power, are small in size and weight, are easier to be put together into a complete equipment and to be aligned, thus reducing the manufacturing cost of the equipment. All of this has led to the wide use of integrated circuits with a large scale of integration, enclosing several functional units in a single package. Also, ICs have improved the key characteristics of equipment. They have brought with them the possibility of building sophisticated pieces of equipment inconceivable with discrete components for economic and technological reasons.

An IC module is a functionally complete subassembly (an amplifier, a frequency converter, a demodulator, and the like) or a part of a receiver, combining several such subassemblies (an r.f. amplifier, an i.f. amplifier and a frequency changer) in a single module.

Radio receivers were first built around hybrid ICs and were set up in traditional circuit configurations. As a rule, they used common-emitter or common-base *npn*-transistor stages, differential stages, and local negative feedback. Because of the low scale of integration, quite a large number of discrete components were needed, such as high-value inductors and capacitors which could not be manufactured by IC technology.

Radio- and intermediate-frequency amplifiers based on hybrid ICs have good temperature and frequency characteristics, a low level noise, and high parameter repeatability. Unfortunately, the low scale of integration stands in the way of cutting down the cost. That is why hybrid IC technology is preferably used in making high-quality receivers intended to operate under adverse conditions and manufactured in limited lots.

Semiconductor ICs are more reliable because they have fewer contact connections and use no discrete components. Also, being smaller in size, they are more robust.

The IC modules most often used in receivers are differential and operational amplifiers whose outputs are coupled to tuned circuits or band-pass filters. The tendency to avoid high-value inductors and capacitors which cannot be made in IC form has spurred the development of coilless active filters which are gradually finding their way into receivers.

Chapter Four

Frequency Converters and Parametric Amplifiers

4.1. General Principles of the Heterodyne Frequency Converter

If a receiver is tuned to an angular frequency ω, its tuned circuits will pass only a part of the voltage spectrum induced by electromagnetic waves in the antenna. As the components of this partial spectrum beat together, they produce a quasi-harmonic voltage waveform

$$v = V(t) \cos [\omega t + \varphi(t)]$$

whose frequency and phase vary in a complex manner.

This waveform is not unlike a modulated signal and in fact it is such a signal (provided the receiver is tuned to the signal frequency and there is no interference present). The amplitude $V(t)$ and the phase $\varphi(t)$ vary at the beat frequencies among which the highest is equal to the maximum difference between the frequencies that make up the signal spectrum, that is, it is equal to the bandwidth B of the section at whose output the extracted spectrum is observed. If the bandwidth is substantially smaller than the frequencies of the incoming wave, these variations are relatively slow, a fact which permits the resultant wave to be treated as a quasi-harmonic one.

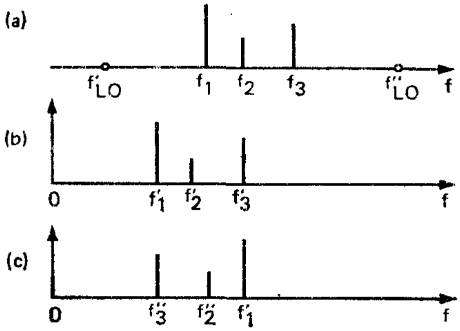


Fig. 4.1

The purpose of a frequency converter is to translate the selected spectrum in frequency without affecting its waveform and, as a consequence, without changing the modulation of the signal which forms or is part of this spectrum. With the aid of a local oscillator operating at angular frequency ω_{LO} and phase φ_{LO}, the signal spectrum is changed to the form

$$v' = KV(t) \cos [\omega \pm k\,\omega_{LO})t + \varphi(t) \pm k\varphi_{LO}]$$

When $k > 1$, we have frequency conversion with local-oscillator harmonics, or frequency conversion of order k.

As an example, consider a signal having a simple spectrum of components f_1, f_2, and f_3, as shown in Fig. 4.1*a*. The local-oscillator frequency f_{LO} may lie below or above these three component frequencies. In the former case, it will be denoted as f'_{LO}, and in the latter, as f''_{LO}. In first-order frequency conversion, when $f_{LO} = f'_{LO}$, the component frequencies f_1, f_2, and f_3 are changed to $f'_1 = f_1 - f'_{LO}$, $f'_2 = f_2 - f'_{LO}$ and $f'_3 = f_3 - f'_{LO}$ as shown in Fig. 4.1*b*. When $f_{LO} = f''_{LO}$, the component frequencies are changed to $f''_1 = f''_{LO} - f_1$, $f''_2 = f''_{LO} - f_2$, and $f''_3 = f''_{LO} - f_3$, as shown in Fig. 4.1*c*. In the latter case, the spectral lines of the converted spectrum are arranged on the frequency axis in a reverse order compared with the cases in Figs. 4.1*a* and *b*. This is what is an inverted spectrum, and the frequency converter which produces it is referred to as an inverting frequency converter.

In a receiver, the spectrum of the incoming wave is translated from any portion of the r.f. spectrum into the passband of the i.f.

amplifier by a circuit which has a variable gain. For parameter control, the circuit contains one or several nonlinear elements, and the local-oscillator voltage is let to act upon them. Preferably, the nonlinearity ought not to markedly affect the spectrum being converted, that is, this action ought not to produce harmonics and intermodulation products. If this requirement is satisfied, the circuit whose parameters are periodically varied under the action of the local oscillator will be a linear one in relation to the spectrum being converted.

The nonlinear elements most commonly used are diodes and transistors. For example, if the local-oscillator voltage $V_{LO} \cos(\omega_{LO} t + \varphi_{LO})$ is impressed on a nonlinear element, its conductance g will vary at frequency f_{LO} and may be represented by a Fourier series as

$$g = g_0 + \sum_{k=1}^{\infty} g_k \cos(k\omega_{LO} t + k\varphi_{LO})$$

If the nonlinear element is acted upon by the voltage of the spectrum being converted, each of its spectral lines, $v_i = V_i \cos(\omega_i t + \varphi_i)$, will give rise to a current, $i = gv_i$. On multiplying together and replacing the products of cosines by functions of sum and difference angles, we get

$$i = g_0 V_i \cos(\omega_i t + \varphi_i) + \sum_{k=1}^{\infty} g_k V_i \cos[(k\omega_{LO} \pm \omega_i) t + k\varphi_{LO} \pm \varphi_i] \tag{4.1}$$

It is seen that to the spectrum of component frequencies f_i are added the shifted spectra of component frequencies $kf_{LO} + f_i$ and $kf_{LO} - f_i$ (or $f_i - kf_{LO}$, if $kf_{LO} < f_i$). Each of the additional spectra has the same structure as the original one. If this is the spectrum of a modulated signal, the latter will remain modulated in the same manner at the new frequency as well. If interference is superimposed on this spectrum, it will be preserved in the converted spectrum, but the interfering frequencies will be different.

Similar results are produced by a nonlinear reactance. If it is an element with a capacitance

$$C = C_0 + \sum_{k=1}^{\infty} C_k \cos(\mathrm{k}\omega_{LO} t + \mathrm{k}\varphi_{LO})$$

the spectral component v_i will give rise in this element to a current

$$i = dq_i/dt$$

where $q = Cv_i$ is the charge. Then,

$$i = C\,(dv_i/dt) + v_i\,(dC/dt)$$

On substituting here for v_i and C, we obtain a sum whose terms contain products of the form

$$v_i\,\omega_i C_0 \sin(\omega_i t + \varphi_i)$$

$$v_i\,\omega_i C_k \cos(k\,\omega_{LO}\, t + k\varphi_{LO}) \sin(\omega_i t + \varphi_i)$$

$$v_i k\,\omega_{LO} C_K \sin(k\,\omega_{LO} t + k\varphi_{LO}) \cos(\omega_i t + \varphi_{LO})$$

It is seen that in this case, too, the current spectrum contains component frequencies $kf_{LO} \pm f_i$ in addition to the original spectrum.

4.2. Ganged Tuning and Tracking

In a receiver with a tuned preselector, the input circuit and the local oscillator are tuned to different frequencies. In the 1920s, the tuned circuits of a superhet would usually be tuned by separate tuning controls, which did not make the tuning an easy job. Later, what is now known as *ganged* (*unicontrol*) *tuning* came to be used. With it, the tuning capacitors of the several r.f. tuned circuits are coupled together mechanically and operated by a single control. This arrangement is shown by the dashed lines in the diagram of Fig. 4.2*a*.

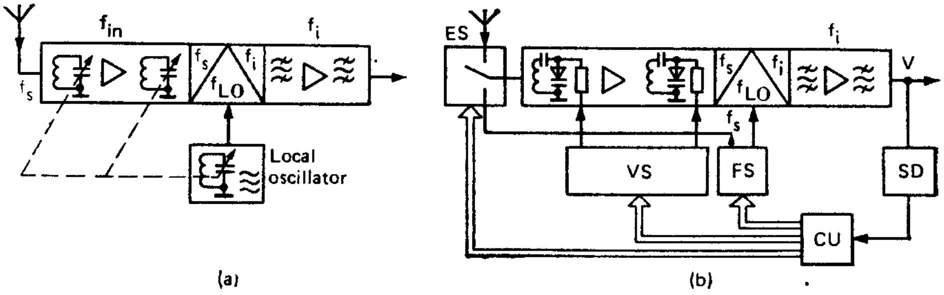


Fig. 4.2

However, this operating convenience poses an additional problem — it becomes necessary to make the various tuned circuits *track*, that is, tune to the same frequency at each setting of the tuning control. With regard to the local oscillator it is required that the local-oscillator tuned circuit(s) differs accurately from the signal by just the amount of the i.f.

Consider the ganged tuning and tracking of tuned circuits in more detail with reference to an 'up-tuned' local oscillator, that is, one whose frequency is higher than the signal because this gives a smaller value for the ratio of maximum to minimum local-oscillator frequency (see Sec. 1.2). In such a case, for the signal frequency f_s to be translated to frequency f_i within the bandwidth of the i.f. amplifier, the local oscillator is set to $f_{\mathrm{LO}} = f_{\mathrm{s}} + f_{\mathrm{i}}$. In turn, in order that the signal could fall within the passband of the preselector, the input circuit should be set to $f_{\mathrm{in}} \approx f_{\mathrm{s}} = f_{\mathrm{LO}} - f_{\mathrm{i}}$. The required maximum to minimum frequency ratio over the band for the input circuit (see Sec. 2.3) is

$$k_{\mathrm{b,in}} = f_{\mathrm{in,max}}/f_{\mathrm{in,min}} = f_{\mathrm{s,max}}/f_{\mathrm{s,min}}$$

For the local oscillator, this ratio is

$$k_{\mathrm{b,LO}} = f_{\mathrm{LO,max}}/f_{\mathrm{LO,min}} = (f_{\mathrm{s,max}} + f_{\mathrm{i}})/(f_{\mathrm{s,min}} + f_{\mathrm{i}})$$

As is seen, $k_{\mathrm{b,LO}} < k_{\mathrm{b,in}}$. In order to reduce the maximum to minimum local-oscillator frequency ratio, one has to add trimmer and padding capacitors to its tuned circuit, as has been described in Sec. 2.3 and shown in Fig. 2.8. With a local oscillator set to operate at a frequency lower than the signal, one has to reduce $k_{\mathrm{b,in}}$ in comparison with $k_{\mathrm{b,LO}}$ in a similar way.

To simplify receiver construction and control, it is usual to make the ganged capacitors identical. The difference in resonance frequency between the tuned circuits is not exactly equal to f_{i} at any value of C_{ckt} in the range from $C_{\mathrm{ckt,min}}$ to $C_{\mathrm{ckt,max}}$. It is seen from Fig. 2.8 that with the characteristic of C_{ckt} specified in advance, the designer is free to vary only three parameters of the tuned circuit having a reduced maximum to minimum frequency ratio, namely the inductance L and the capacitances C_1 and C_2. The three parameters should be chosen such that the desired value of frequency is obtained at three settings of C_{ckt}. Two of them are chosen near $C_{\mathrm{ckt,min}}$ and $C_{\mathrm{ckt,max}}$, respectively, and the third, in between them. This gives what is known as *three-point tracking*, that is, alignment is perfect at three points in the band. At any other value of C_{ckt}, the difference frequency is not equal to f_{i}; rather, it takes on a slightly different value, f_{i}'.

It is vitally essential for the local oscillator to be exactly tuned — failure to do so would shift the spectrum of the received signal outside the passband of the i.f. section. Therefore, an error in tracking leads to an inaccurate tuning of the preselector, that is, its resonance frequency is no longer equal to the signal frequency. If this is not to impair the quality of reception, the bandwidth of the preselector is spread on either side of the resonance frequency by an amount, $\delta f_{0,\mathrm{pr}} = | f_{\mathrm{i}} - f_{\mathrm{i}}' |$, equal to the *tracking error*. The value of C_{ckt} at which perfect tracking is obtained is chosen such that $\delta f_{\mathrm{track}}$ is a minimum. Then an increase in bandwidth will not materially impair the selectivity of the preselector.

When the maximum to minimum frequency ratio is relatively small, satisfactory tracking can be obtained at two points in the band, so one may limit oneself to only one capacitor, C_1 or C_2.

Now that mechanical tuning is giving way to electronic tuning and an ever wider use is made of automatic control devices based on microelectronics and digital techniques, there is a trend towards the individual control of tuned circuits. In approximate form, such a tuning control scheme is shown in Fig. 4.2*b*. Here, *CU* is a control unit usually consisting of a microprocessor, a memory, and programmers additionally linked to pushbutton or any other manual controls. The required local-oscillator frequency f_{LO} is supplied by a digital frequency synthesizer, *FS*. The latter also generates the signal frequency f_s to which the receiver should be tuned. For this purpose, f_i is subtracted from or added to f_{LO}. There is also a voltage synthesizer, *VS*, which feeds suitable voltages to the varactors in the preselector in order to tune it to the desired frequency.

To begin with, the control unit operates to move the electronic switch, *ES*, to its lowermost position, and a voltage at f_s is applied to the receiver input. At the same time, the voltage synthesizer feeds a staircase tuning voltage to the varactor in the input tuned circuit. As resonance is approached, a voltage *V* is generated at the i.f. amplifier output; it is detected by the signal detector, *SD*, and applied to the control unit. When this voltage reaches its maximum, which corresponds to exact tuning, adjustment of the input tuned circuit ceases, and the next tuned circuit is aligned in a similar way. After the preselector has been aligned, the electronic switch is moved to its uppermost position, and this makes the receiver operative.

4.3. Frequency Conversion Spurious Products

Figure 4.3 shows a frequency converter which consists of a mixer *Mxr*, a local oscillator *LO*, an input filter *Filt*1 and an output filter, *Filt*2, tuned to the intermediate frequency, f_i. Let us apply a sine-wave voltage v_{in} at frequency f to the mixer. Also let us vary f and see how the output voltage is varying.

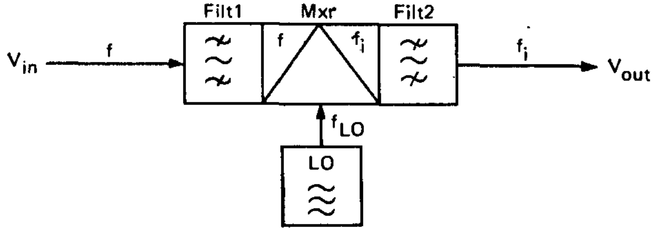


Fig. 4.3

When f is equal to f_i, that is when the first current component, Eq. (4.1), has the same frequency as that of *Filt*2, the output voltage

v_{out} will be a maximum. As the frequency is varied, the output voltage will decrease according to the frequency response of *Filt*2.

Considering the second term in Eq. (4.1) for $k = 1$, we find that resonance will also occur at $f_{LO} - f = f_i$ and $f - f_{LO} = f_1$, that is, at frequencies $f = f_{LO} + f_1$ and $f = f_{LO} - f_1$.

Similarly, considering the case for $k = 2$, we can see that resonance occurs at frequencies $2f_{LO} - f_1$, $2f_{LO} + f_1$, and so on. Thus, the frequency response have several peaks (*1*, *2*, *3*, . . ., as shown in Fig. 4.4). As a rule, the higher the conversion order, the lower the amplitude of the respective current components.

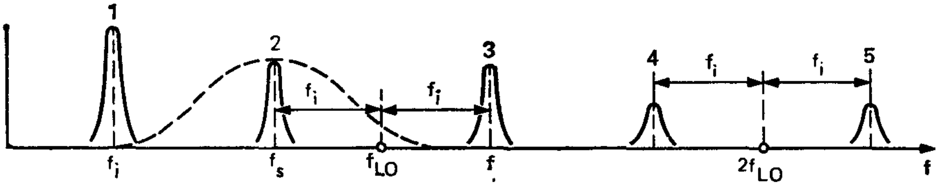


Fig. 4.4

Since the purpose of a frequency converter is to translate the signal spectrum in frequency, peak *1* is not utilized because it is due to the passage of the received signal without frequency conversion. On the other hand, if the interfering signals reaching the frequency converter are at the intermediate frequency f_1, they will fall within the filter passband and, being superimposed on the received signal, they will adversely affect its normal reception.

Also falling within the passband of *Filt*2 will be the conversion products at frequencies where additional resonances (*2*, *3* . . .) take place. One of these responses or channels is the desired one and all the other are spurious responses (a general outline of spurious responses has been given in Chap. 1). Suppose that the desired signal at frequency f_s is properly received within the frequency band where peak *2* occurs. Then response *2* may be called the desired channel, and response *3*, a spurious channel. Relative to f_{LO} it is a mirror image of the desired channel, and quite aptly it is called the *image response* or simply the *image*. That is why in Fig. 4.4 the respective frequency is labelled as f_{im}. As an alternative, the frequency band where resonance *3* occurs may be chosen as the desired one, in which case the channel corresponding to resonance *2* will be the image. In each case, the desired signal and the image differ by $2f_1$ in frequency. Furthermore, the frequency bands around resonance *4* or *5* may be chosen as the desired channel, but this is seldom done because second-order frequency conversion is lower in efficiency. Still, if this is the case, the frequency bands within which resonances *2* and *3* occur would be spurious channels of reception.

Since all frequencies at which resonances *1*, *2*, *3*, etc. occur would produce a voltage within the passband of *Filt*2 (see Fig. 4.3) and cannot be suppressed beyond that filter, measures are usually taken to prevent any unwanted frequencies from reaching the frequency converter. They should be suppressed in *Filt*1 (see Fig. 4.3).

Because the i.f. is fixed, any spurious response at that frequency, f_1, is not difficult to suppress. This purpose can be served by merely placing ahead of the frequency converter a rejector circuit tuned to that frequency. If a receiver is intended for operation at a fixed frequency, the image frequency will also be a fixed one, and image interference can then be suppressed by an untunable filter. This problem is more difficult to tackle in a tunable receiver because the image frequency will be then a varying quantity. To facilitate image rejection, one has to choose an intermediate frequency which is not too low. As is seen from Fig. 4.4, the image and the signal are moved so far apart that, even though *Filt*1 has a relatively wide passband (its characteristic is shown in the figure by the dashed curve), image rejection will be considerable.

As has been noted, the local oscillator may be set at a frequency which is higher or lower than the signal frequency by an amount equal to the i.f. In the former case, the image frequency is higher than f_s by twice the i.f. In the latter case, it is lower than the signal by twice the i.f. Therefore, the image frequencies with the local oscillator tuned 'up' and 'down' are spaced four time the i.f. apart. The frequency response of the tuned circuit far away from resonance is unsymmetrical, running steeper to the left and more gradually to the right of the resonance frequency. That is why the image selectance is usually greater with the local oscillator set to a frequency lower than the signal as compared with the case where it set to a frequency higher than the signal.

4.4. The Image-Compensated Frequency Converter

Sometimes it may so happen that the input filter (*Filt*1 in Fig. 4.3) cannot give the desired image rejection whereas an increase in the i.f. is undesirable. In a situation like that, resort is made to an image-compensated frequency converter set up as shown in Fig. 4.5. The received signal has the same phase in the two arms of the frequency converter and is therefore doubled in magnitude in the common circuit whereas the image frequencies are in anti-phase and cancel out.

The voltage

$$v_{\text{LO}} = V_{\text{LO}} \cos\left(\omega_{\text{LO}} t + \varphi_{\text{LO}}\right)$$

generated by the local oscillator, *LO*, is fed to one mixer, Mxr_1, with a phase shift of $+\pi/4$ produced by a phase shifter, PS_1, and to the other mixer, Mxr_2, with a phase shift of —π/4 produced by another phase shifter, PS_2. The voltage at the intermediate frequency, f_i, appearing at the output of the first mixer is selected by a filter, $Filt_2$, and additionally shifted in phase by +π/4 by a third phase shifter, PS_3. The voltage appearing at the output of the second mixer, Mxr_2, is additionally shifted by — π/4 in phase by a fourth phase shifter, PS_4. The two arms of the frequency converter are arranged to have the same transmission gain.

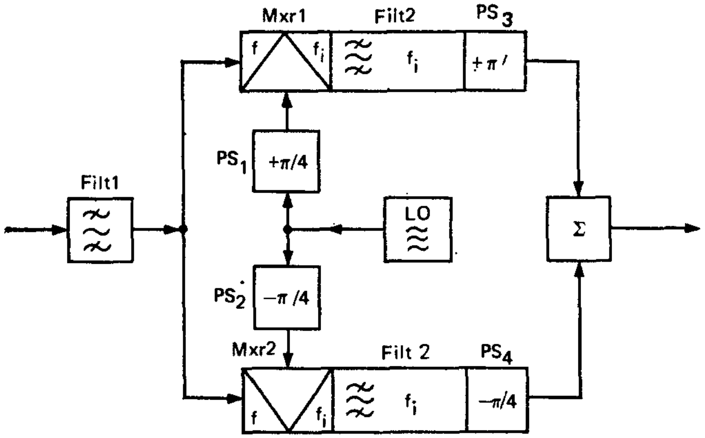


Fig. 4.5

Suppose that the two mixers Mxr_1 and Mxr_2, are fed voltages at the signal frequency

$$v_s = V_s \cos(\omega_s t + \varphi_s)$$

and at the image frequency

$$v_{im} = V_{im} \cos(\omega_{im} t + \varphi_{im})$$

such that $f_s - f_{LO} = f_i$, $f_{LO} - f_{im} = f_{i2}$, and $f_{i1} \approx f_{i2} \approx f_i$.

Past the first mixer, the signal and image voltages are

$$v_{s1} = V_s K_{mxr} \cos[(\omega_s - \omega_{LO})t + \varphi_s - \varphi_{LO} + \pi/4)]$$

and

$$V_{im1} = V_{im} K_{mxr} \cos[(\omega_{LO} - \omega_{im})t - \varphi_{im} + (\varphi_{LO} + \pi/4)]$$

and past the second mixer they are

$$v_{s2} = V_s K_{mxr} \cos[(\omega_s - \omega_{LO})t + \varphi_s - (\varphi_{LO} - \pi/4)$$

$$v_{im2} = V_{im} K_{mxr} \cos[(\omega_{LO} - \omega_{im})t - \varphi_{im} + (\varphi_{LO} - \pi/4)]$$

Here, K_{mxr} is the transmission gain of the two mixers, Mxr_1 and Mxr_2, and the two filters, $Filt'_2$ and $Filt''_3$, taken together.

Past the phase-shifters PS_3 and PS_4 we have in the upper arm

$$v'_{s1} = V_s K_{mxr} K_{PS} \cos[(\omega_s - \omega_{LO})t + \varphi_s - \varphi_{LO}]$$

$$v'_{im1} = V_{im} K_{mxr} K_{PS} \cos[(\omega_{LO} - \omega_{im})\, t - \varphi_s + \varphi_{LO} + \pi/2]$$

and in the lower arm

$$v'_{s2} = V_s K_{mxr} K_{PS} \cos[(\omega_s - \omega_{LO})t + \varphi_s - \varphi_{LO}]$$

$$v'_{im2} = V_{im} K_{mxr} K_{PS} \cos[(\omega_{LO} - \omega_{im})t - \varphi_s + \varphi_{LO} - \pi/2]$$

Here, K_{PS} is the transmission gain of the phase shifters, PS_3 and PS_4.

It is seen from the above equations that v'_{s1} and v'_{s2} are in phase, whereas v'_{im1} and v'_{im2} are in anti-phase. For this reason, in the common circuit the signal amplitude is doubled, being $2V_s K_{mxr} K_{PS}$, whereas the image voltages cancel out.

4.5. Double Frequency Conversion

As previously mentioned (see Sec. 4.3 and Fig. 4.4), the image is progressively easier to eliminate as the image and the signal are spaced farther apart, that is, as a progressively higher i.f. is chosen. On the other hand, better adjacent-channel attenuation and stable gain are easier to obtain with a lower i.f. This conflict can be resolved by using double or even triple frequency conversion. The circuit of a double frequency converter is shown in Fig. 4.6*a*.

If the signal frequency is to be translated to a frequency which is lower by a factor of several hundred or more, this is done in several consecutive steps rather than at once. To begin with, f_s is heterodyned by a mixer, Mxr_1, and a local oscillator, LO_1, to the first intermediate frequency, f_{i1}, which is one-tenth to one-twentieth of the signal frequency. The image frequency differs from the signal frequency by 10 to 20%, a fact which permits the first filter, *Filt*1, to suppress it appreciably. The first i.f. is extracted by a second filter, *Filt*2, and is further translated to the required value, f_{i2}, in a second frequency converter made up of Mxr_2, LO_2, and a filter, *Filt*3, tuned to the second i.f.

A disadvantage of double conversion is the production of a second image response. To demonstrate, if the second conversion is effected such that

$$f_{i2} = f_{LO2} - f_{i1}$$

the same frequency will be produced if the voltage appearing at the output of the second filter. *Filt*2, has a frequency $f' = f_{i1} + 2f_{i2}$, because it will be heterodyned in the following manner:

$$f' - f_{LO2} = (f_{i1} + 2f_{i2}) - (f_{i1} + f_{i2}) = f_{i2}$$

and it will likewise be selected by *Filt*3.

The second image frequency differs from the signal frequency by $2f_{i2}$, which is a factor of several tens smaller than the difference in the case of the first image frequency. This interference cannot markedly be suppressed in *Filt1* and must be rejected by' *Filt2*, which is the primary purpose of the latter.

There is a widely used alternative double conversion scheme; it is shown in Fig. 4.6*b*. Although outwardly it is nearly the same as the previous one, it offers a number of important advantages. In this case, the first i.f., f_{i1}, is chosen to be higher rather than lower than the maximum value of the signal frequency, f_s. Thus, the incoming signal is raised rather than lowered in frequency. Such a receiver is called the infradyne. The first i.f. is then heterodyned in the second mixer, *Mxr2*, to a second i.f., f_{i2}, which is selected by a filter, *Filt3*, whereas the second image frequency is rejected by *Filt2*.

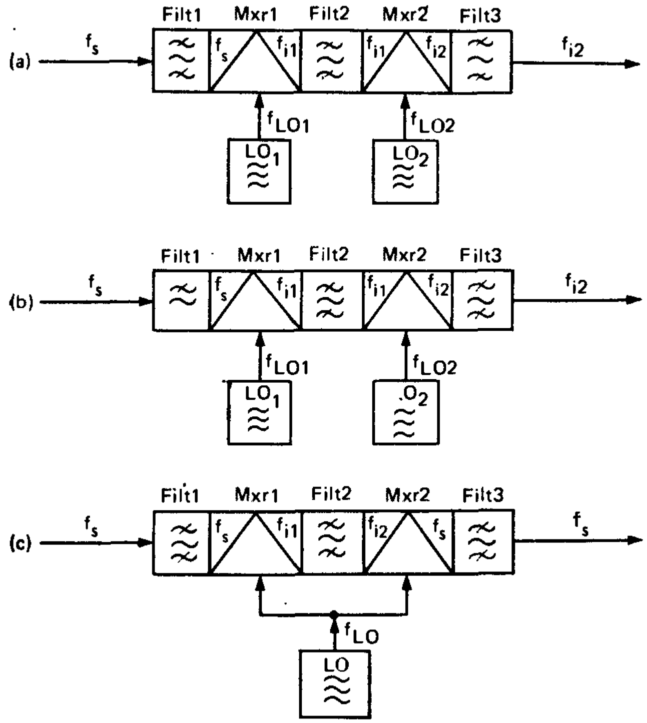


Fig. 4.6

Since f_{i1} is high, the second mixer alone may prove inadequate to lower it to the desired value for the reasons set forth above. Because of this, a third conversion (not shown in Fig. 4.6*b*) may be required in an infradyne. If so, a third image will be produced; it can be suppressed by *Filt3*.

One of the advantages offered by an infradyne is the simplified construction of *Filt*1. In a receiver tunable over a wide frequency range this filter is undesirable because it calls for continuous tuning in a band and coil-switching for band selection, while a mechanical band selector switch is far from simple to make, cannot be miniaturized, and is of low reliability. Indeed, switch-contact wear is often the cause of receiver failure.

When $f_{i1} > f_{s,max}$, the spurious i.f. response falls outside the frequency range of the receiver (see Fig. 4.4). The image likewise occurs beyond the upper limiting frequency of the range because with $f_{LO1} = f_s + f_{i1}$ it will range from $f_{s,min} + 2f_i$ to $f_{s,max} + 2f_i$. This permits the use of a fixed-tuned low-pass filter for *Filt*1, which will pass all frequencies below $f_{s,max}$ to the input of the first mixer.

Another advantage of the infradyne is a substantially reduced maximum to minimum frequency ratio of the first local oscillator. In a down-conversion receiver with $f_{LO1} = f_s + f_{i1}$, this ratio is

$$K_{b,LO} = (f_{s,max} + f_{i1})/(f_{s,min} + f_{i1})$$

or, to state this differently,

$$K_{b,LO} = (K_{b,s} + \varkappa_{dc})/(1 + \varkappa_{dc})$$

where

$$K_{b,s} = f_{s,max}/f_{s,min}$$

is the maximum to minimum receiver frequency ratio, and

$$\varkappa_{dc} = f_{i1}/f_{s,min}$$

is the down-conversion ratio of the first mixer. As has been shown in Sec. 4.2, this ratio is less than unity, therefore $K_{b,LO}$ differs only slightly from $K_{b,s}$.

For an infradyne, $K_{d,LO}$ may be defined as

$$K_{d,LO} = (\varkappa_{uc} + 1)/(\varkappa_{uc} + 1/K_{b,s})$$

where

$$\varkappa_{uc} = f_{i1}/f_{s,max}$$

is the up-conversion ratio of the mixer. As this ratio increases, $K_{b,LO}$ tends to unity.

If, for instance, $f_{s,min} = 0.3$ MHz and $f_{s,max} = 30$ MHz, then $K_{b,s} = 100$. For proper coverage, this frequency range would have to be divided into 4 switch-selectable bands. A similar provision of band-switching would have to be made in the local oscillator as well. In an infradyne receiver, if one chooses for example, $f_{i1} = 60$ MHz, no switching is required in the r.f. section; it will suffice to have a low-pass filter (see Fig. 4.6*b*). Then $\varkappa_{uc} = 2$ and $K_{b,LO} = (2 + 1)/(2 + 0.01) = 1.5$. With the maximum-to-minimum frequency ratio that small, no band selector switch is necessary for the local oscillator.

Figure 4.6*c* shows still another double conversion scheme. In a down-converting mixer, *Mxr*1, the signal frequency f_s is heterodyned to f_1. Following that a second mixer, *Mxr*2, up-converts it so as to restore the initial frequency f_s. Suppose that the voltage applied to the device via a filter, *Filt*1, which suppresses the image interference, is of the form

$$v_s = V_s \cos(\omega_s t + \varphi_s)$$

and that the local oscillator generates a voltage

$$v_{LO} = V_{LO} \cos(\omega_{LO}\, t + \varphi_{LO})$$

The voltage appearing at the output of *Mxr*1 will then have the form

$$v_1 = K_{mxr1} V_s \cos(\omega_1 t + \varphi_1)$$

such that

$$\omega_1 = \omega_{LO} - \omega_s$$

$$\varphi_1 = \varphi_{LO} - \varphi_s$$

and K_{mxr1} is the transmission gain of *Mxr*1. This voltage is fed via *Filt*2 to *Mxr*2. The voltage appearing at the output of *Filt*3 is

$$v_{out} = K_{mxr1} K_{mxr2} V_s \cos[(\omega_{LO} - \omega_1)t + \varphi_s - \varphi_1]$$

On substituting for ω_1 and φ_1, we obtain

$$v_{out} = K_{mxr1} K_{mxr2} V_s \cos(\omega_s t + \varphi_s)$$

As is seen, the output voltage has the same frequency and is in phase with the input voltage, if we neglect the phase shifts likely to occur in the three filters.

The scheme just examined has three important properties, namely:

— the bandwidth and selectivity are determined by *Filt*2. If $f_1 \ll f_s$, the bandwidth may be made narrow and the selectivity high. In consequence, the circuit may operate as a narrowband filter which is difficult to implement at f_s;

— the phase of the output signal is independent of that of the local-oscillator voltage and the instability of the latter does not affect the frequency of the output signal because the shifts in frequency and phase in *Mxr*1 and *Mxr*2 are mutually opposite and cancel out.

The above method of filtering is advantageous in cases where it is required to obtain high frequency and phase stability. It is employed in several modifications in receivers intended for various services.

4.6. Frequency Converter Types

The nonlinear elements used in frequency converters are mostly transistors and diodes. The key difference between transistor and diode frequency converters consists in that a transistor in a nonreciprocal device, which means that the input voltage affects the output current differently from how the output voltage affects the input current. In a diode, the current is common to both the input and output and the two voltages affect this current in the same way, which is another way of saying that circuits containing diodes fall in the class of reciprocal circuits.

Given certain conditions, a transistor, a tunnel diode and a varactor (a voltage-variable capacitor) are able to amplify radio signals. Therefore they can be used to build active frequency converters which, in addition to conversion, can provide amplification. A rectifying diode attenuates rather than amplifies the signal being converted, which means that it acts as a passive converter.

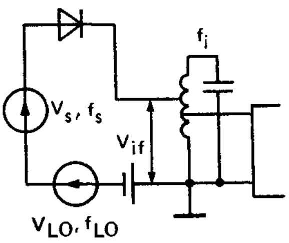


Fig. 4.7

An amplifying device can be used to generate oscillations. In the frequency converters shown in Figs. 4.3, 4.5 and 4.6, the local oscillators can be built around a transistor, a negative-resistance diode (a negatron), a vacuum tube, or any other amplifying device. In active converters, the electron device may combine the functions of both a mixer and a local oscillator, in which case it will be referred to as an *autodyne* or *self-oscillating frequency converter*. Since the optimal operating conditions for the generation of oscillations and frequency conversion are different, it is more common to use frequency converters with a separate local oscillator.

When choosing operating conditions for the electron devices of a frequency converter, one seeks to ensure a maximum transmission gain, linear signal conversion, a minimal level of internal noise, a minimal level for spurious responses which might interfere with the reception of the desired signal, and a minimal amount of coupling between the r.f. circuits and the local oscillator. With tight coupling between them, the receiver is difficult to align and the oscillations generated by the local oscillator might be radiated by the antenna, thus causing interference to other receivers.

In the diode frequency converter of Fig. 4.7, the signal source and the local oscillator are connected in the diode circuit which also shapes the i.f. voltage. This diagram does not show that the source of the signal voltage being heterodyned, V_s, is the input circuit or the r.f. amplifier; this source is traversed by the electron-device current which has a complex spectrum. Since this current is other than sinusoidal in waveform, the voltage across the input circuit may likewise be nonsinusoidal. However, the input circuit contains a resonant circuit tuned to the signal frequency f_s, across which the voltage drop is produced practically by the current at the fundamental frequency alone. For this reason it is legitimate to take v_s to be a quasi-harmonic, or sinusoidal, one with an amplitude and phase slowly varying in step with the signal modulation.

The local oscillator likewise contains a resonant circuit tuned to $f_{\rm LO}$; therefore, $v_{\rm LO}$ may similarly be regarded as sinusoidal.

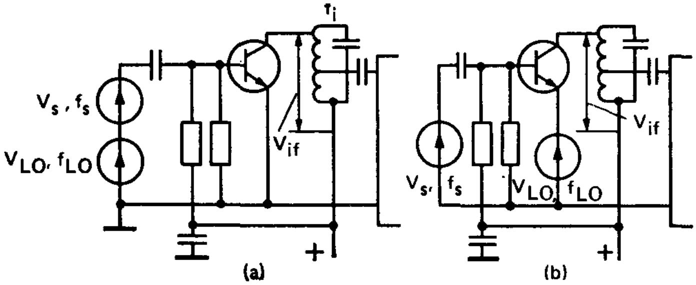


Fig. 4.8

By the same token, the output voltage of the mixer, $v_{\rm mxr}$, developed across the tuned circuit at resonance frequency $f_{\rm i}$ may be regarded as quasiharmonic.

Figure 4.8*a* and *b* show two frequency converters using a nonreciprocal electron device which is a transistor in this case. A similar arrangement can be used for a frequency converter based on a FET or

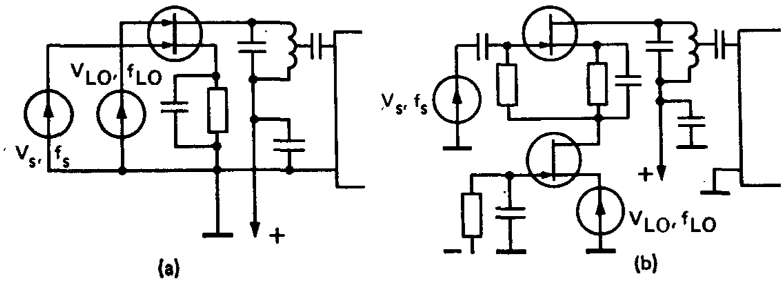


Fig. 4.9

a vacuum tube. The signal source and the local oscillator are seen to be connected between the base and emitter (see Fig. 4.8*a*). In the circuit of Fig. 4.8*b*, the coupling between the mixer and the local oscillator is less tight. For a still looser coupling, the signal and local-oscillator voltages may be applied to different electrodes, as shown in Fig. 4.9.

In the circuit of Fig. 4.9*a*, the signal and local-oscillator voltages are applied to different gates of a double-gate FET. In the frequency converter of Fig. 4.9*b*, these two voltages are applied to the gates of two series-connected FETs. The local-oscillator voltage may alternatively be applied to the source electrode rather than the gate of the lower FET.

Figure 4.10 gives two examples of an autodyne frequency converter. The local-oscillator current is fed from the collector to the feedback circuit of the local oscillator which is a transformer-coupled circuit in Fig. 4.10*a*, and a tapped-coil-coupled circuit in Fig. 4.10*b*.

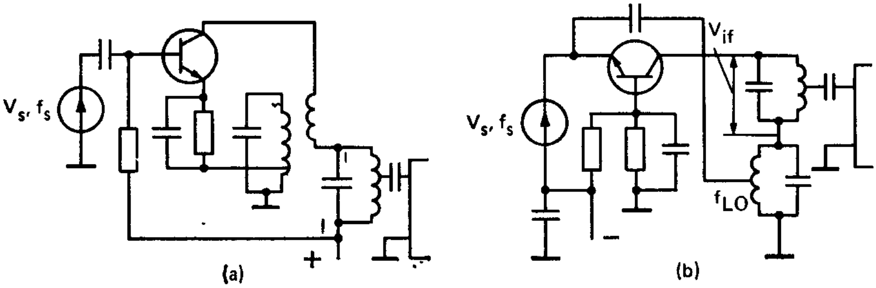


Fig. 4.10

4.7. The Theory of Frequency Conversion by a Nonreciprocal Electron Device

A generalized circuit diagram of a nonreciprocal frequency converter is shown in Fig. 4.11 where *ED* is an electron device or an assembly of several electron devices (such as an IC) along with auxiliary circuits, acting as a mixer. The mixer has two inputs, one for the signal voltage, v_s, and the other for the local-oscillator voltage, v_{LO}, and one output for the i.f. voltage, v_{if}. In all of the three circuits, direct voltages may exist; they come from the respective power supplies and maintain the required operating conditions for the electron device(s).

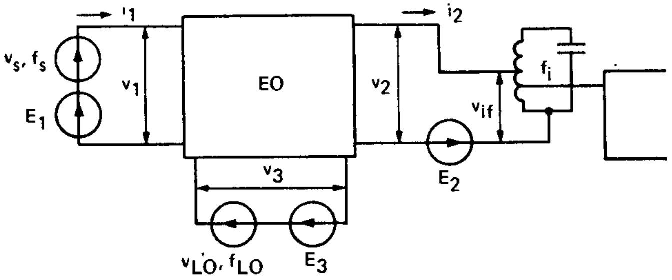


Fig. 4.11

Frequency conversion in a receiver is specific in that, as has been noted, the signal voltage being heterodyned is relatively low, being a small fraction of the local-oscillator voltage. Therefore, the signal applied to the electron device does not affect the current and power drawn from the local oscillator. For this reason it is legitimate to calculate the current and power drawn by the local oscillator on the assumption that no signal voltage v_s exists at the input and that no i.f. voltage exists at the output.

The volt-ampere characteristic of a mixer regarded as the load of the associated local oscillator can be approximated by a power polynomial. With a sinusoidal local-oscillator voltage, the current in the circuit contains the fundamental and harmonics. The input impedance of the mixer as the local-oscillator load can be found on dividing the amplitude of the local-oscillator voltage by that of the current at the fundamental frequency.

Since v_s and, as a consequence, v_{if} are relatively low, frequency conversion may be analysed by invoking a simple method from the theory of nonlinear circuits: the mixer currents as functions of the applied voltages are expanded into Taylor series in powers of these low voltages, and the terms containing high powers of the infinitesimals are discarded.

In the general case, when analysing a frequency converter, it is important to consider the internal capacitances of the electron devices. They vary with the applied voltages, that is, are nonlinear, and affect therefore the process of frequency conversion (see Sec. 4.1). The need to take into account the complex nonlinearity complicates frequency-conversion theory. If the signal frequency lies well below the limiting frequency of the electron device, the reactances will only slightly affect the performance of the frequency converter. Therefore, the parameters of the electron devices used in nonreciprocal frequency converters may to a first approximation be deemed real and frequency-independent.

Since the external circuits of a frequency converter contain resonant circuits tuned to the respective frequencies (f_s, f_i, f_{LO}), it is legitimate to assume that the currents at any other frequencies (harmonics, intermodulation products) do not produce marked voltages across these circuits.

The input and output currents of the frequency converter shown in Fig. 4.11 may be written as the following functions:

$$i_1 = f_1 (v_{LO}, v_s, f_{if}, E_1, E_2, E_3)$$

$$i_2 = f_2 (v_{LO}, v_s, v_{if}, E_1, E_2, E_3)$$

Let us expand the two functions into Taylor series and, assuming that the received signal is fairly weak, limit ourselves to the expansion terms containing v_s and v_{if} to the first power

$$
\begin{aligned}
i_1 &= f_1(v_{LO},\ E_1,\ E_2,\ E_3) + \frac{\partial f_1(v_{LO},\ E_1\ E_2,\ E_3)}{\partial v_1} v_s \\
&\qquad + \frac{\partial f_1(v_{LO},\ E_1,\ E_2,\ E_3)}{\partial v_2} v_{if} + \dots \\
i_2 &= f_2(v_{LO},\ E_1,\ E_2,\ E_3) + \frac{\partial f_2(v_{LO},\ E_1,\ E_2,\ E_3)}{\partial v_1} v_s \\
&\qquad + \frac{\partial f_2(v_{LO},\ E_1,\ E_2,\ E_3)}{\partial v_2} v_{if} + \dots
\end{aligned}
\tag{4.2}
$$

In Eqs. (4.2), the terms $f_1\,(v_{LO},\ E_1,\ \dots)$ and $f_2\,(v_{LO},\ E_1\ \dots)$ are the currents i_1 and i_2 in the absence of the signal being converted, when the mixer is only fed the local-oscillator and supply voltages. Let us denote them as i_{1LO} and i_{2LO}. They do not contain components at frequencies f_s and f_i, and their effect on frequency conversion may often be neglected.

The derivative $\partial f_1\,(v_{LO},\ E_1,\ \dots)/\partial v_1$ is the differential input conductance of the mixer, determined when it is only fed the local-oscillator voltage. Let us denote it as g_{in}. Referring to the volt-ampere characteristics, we can construct a plot of g_{in} as a function of v_3. Then, using an approximation of this function or its graphic representation, we can respectively calculate or construct a plot of g_{in} as a function of local-oscillator frequency. Because of nonlinearity, this function will be nonsinusoidal and may be represented as a Fourier series

$$g_{in} = g_{in,0} + \sum_{k=1}^{\infty} g_{in,k} \cos k\omega_{LO} t$$

The derivative $\partial f_1\,(v_{LO},\ E_1,\ \dots)/\partial v_2$ is the differential internal-feedback conductance, g_{fb}, under the same conditions. Similarly to g_{in}, it varies at the local-oscillator frequency and may be written as a Fourier series

$$g_{fb} = g_{fb,0} + \sum_{k=1}^{\infty} q_{fb,k} \cos k\omega_{LO} t$$

The derivative $\partial f_2\,(v_{LO},\ E_1,\ \dots)/\partial v_1$ is the slope, S, of the curve relating output current to input voltage, or the transconductance of the electron device involved*. For the same reasons,

$$S = S_0 + \sum_{k=1}^{\infty} S_k \cos k\omega_{LO} t$$

The derivative $\partial f_2\,(v_{LO},\ E_1,\ \dots)/\partial v_2$ is the differential output conductance $g_{out,c}$ of the frequency converter, which may likewise

* As an alternative, many authors designate this quantity as g_m, as done inchap. 3. — *Translator's note.*

be written as a Fourier series

$$g_{\mathrm{out,c}} = g_{\mathrm{out,c,0}} + \sum_{k=1}^{\infty} g_{\mathrm{out,c,k}} (\cos k\omega_{\mathrm{LO}} t)$$

Signal modulation may be assumed to be relatively slow, and the voltages v_{s} and v_{if} may be treated as quasi-harmonic, (see Sec. 4.1), that is,

$$v_{\mathrm{s}} = V_{\mathrm{s}} \cos (\omega_{\mathrm{s}} t + \varphi_{\mathrm{s}})$$

and

$$v_{\mathrm{if}} = V_{\mathrm{if}} \cos (\omega_{\mathrm{if}} t + \varphi_{\mathrm{if}})$$

where the frequencies ω_{s} and ω_{if} and the phases φ_{s} and φ_{if} may be deemed constant. When considering amplitude and angle modulation, these processes may be considered to be slow.

On substituting for v_{s} and v_{if} in Eq. (4.2) and replacing the product of cosines by the cosines of sum and difference arguments, we get

$$\begin{aligned}
i_1 &= i_{1,\mathrm{LO}} + g_{\mathrm{in,0}} V_{\mathrm{s}} \cos (\omega_{\mathrm{s}} t + \varphi_{\mathrm{s}}) \\
&+ \sum_{k=1}^{\infty} 0.5 g_{\mathrm{out}} V_{\mathrm{s}} [(k\omega_{\mathrm{LO}} \pm \omega_{\mathrm{s}})\, t \pm \varphi_{\mathrm{s}}] \\
&+ g_{\mathrm{fb,0}} V_{\mathrm{if}} \cos (\omega_{\mathrm{if}}\, t + \varphi_{\mathrm{if}}) \\
&+ \sum_{k=1}^{\infty} 0.5 g_{\mathrm{fb}} V_{\mathrm{if}} \cos [(k\omega_{\mathrm{LO}} \pm \omega_{\mathrm{if}})\, t \pm \varphi_{\mathrm{if}}] \\
i_2 &= i_{2,\mathrm{LO}} + S_0 V_{\mathrm{s}} \cos (\omega_{\mathrm{s}} t + \varphi_{\mathrm{s}}) \\
&+ \sum_{k=1}^{\infty} 0.5 S_{\mathrm{k}} V_{\mathrm{s}} \cos [(k\omega_{\mathrm{LO}} \pm \varphi_{\mathrm{s}})] \\
&+ g_{\mathrm{out,0}} V_{\mathrm{if}} \cos (\omega_{\mathrm{if}} + \varphi_{\mathrm{if}}) \\
&+ \sum_{k=1}^{\infty} 0.5 g_{\mathrm{out,c,k}} V_{\mathrm{if}} \cos [k\omega_{\mathrm{LO}} \pm \omega_{\mathrm{if}})\, t \pm \varphi_{\mathrm{if}}]
\end{aligned}$$

Let us find the amplitude of I_{if}, that is, the peak value of current at the intermediate frequency f_{i} in the spectrum of current i_2, and that of I_{s}, that is, the peak value of current at the signal frequency f_{s} in the spectrum of current i_1. Most often,

$$f_{\mathrm{i}} = kf_{\mathrm{LO}} - f_{\mathrm{s}}$$

Let it be case "a". More seldom,

$$f_{\mathrm{i}} = f_{\mathrm{s}} - kf_{\mathrm{LO}}$$

Let it be case "b". Accordingly,

$$f_{\mathrm{s}} = kf_{\mathrm{LO}} - f_{\mathrm{i}}$$

or

$$f_s = kf_{\mathrm{LO}} + f_i$$

As has been noted in Sec. 4.1, use is made of first-order conversion when $k = 1$. Then the sought components of current are

$$i_{\mathrm{if}} = g_{\mathrm{out,c,0}} V_{\mathrm{if}} \cos(\omega_{\mathrm{if}} t + \varphi_{\mathrm{if}}) + 0.5\, S_{\mathrm{k}} V_{\mathrm{s}} \cos(\omega_{\mathrm{if}} t \pm \varphi_{\mathrm{s}})$$
$$i_s = g_{\mathrm{in,0}} V_{\mathrm{s}} \cos(\omega_{\mathrm{s}} t + \varphi_{\mathrm{s}}) + 0.5\, g_{\mathrm{fb,k}} V_{\mathrm{if}} \cos(\omega_{\mathrm{s}} t \pm \varphi_{\mathrm{if}})$$

For case "a", the complex amplitudes of currents are

$$\left.\begin{aligned} \dot{I}_{\mathrm{if}} &= g_{\mathrm{out,c,0}} \dot{V}_{\mathrm{if}} + 0.5 S_{\mathrm{k}} V_{\mathrm{s}}^{*} \\ \dot{I}_{\mathrm{s}} &= g_{\mathrm{in,0}} \dot{V}_{\mathrm{s}} + 0.5 g_{\mathrm{fb,k}} V_{\mathrm{if}}^{*} \end{aligned}\right\} \qquad (4.3a)$$

and for case "b",

$$\left.\begin{aligned} \dot{I}_{\mathrm{if}} &= g_{\mathrm{out,0}} V_{\mathrm{if}} + 0.5 S_{\mathrm{k}} \dot{V}_{\mathrm{s}} \\ \dot{I}_{g} &= {}_{\mathrm{out,0}} \dot{V}_{\mathrm{s}} + 0.5 g_{\mathrm{fb,k}} \dot{V}_{\mathrm{if}} \end{aligned}\right\} \qquad (4.3b)$$

In case "a", V_{s}^{*} and V_{if}^{*} are complex conjugates; they are invoked to stress the fact that the phase angles in the second terms on the right-hand sides of Eqs. 4.3*a* take signs opposite to those of the phase angles of $\dot{V}_{\mathrm{s}}$ and $\dot{V}_{\mathrm{if}}$. This takes place in the inverting frequency converter mentioned in Sec. 4.1.

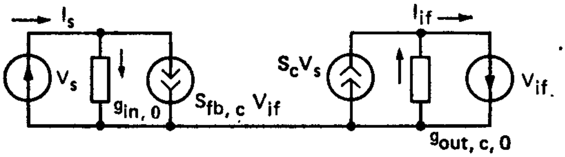


Fig. 4.12

In case "b", the frequency converter does not invert the signal spectrum. In an infradyne receiver, the frequency converter is likewise a noninverting one, if $f_i = f_s + f_{\mathrm{LO}}$. Since $|\,\dot{V}_{\mathrm{s}}\,| = \dot{V}^{*}_{\mathrm{s}}\,|$, the difference between Eqs. (4.3*a*) and (4.3*b*) is not related to the difference between current and voltage amplitudes.

The quantity $S_{\mathrm{c}} = 0.5\, S_{\mathrm{k}}$ is the conversion transconductance and the quantity $S_{\mathrm{fb,c}} = 0.5\, S_{\mathrm{fb,k}}$ is the internal feedback conversion transconductance*.

Proceeding from Eqs. (4.3), one is in a position to set up a model for the frequency converter as shown in Fig. 4.12. These equations

* Many authors use the term "conversion conductance" instead of "conversion transconductance".— *Translator's note.*

may be re-cast as

$$\left.\begin{aligned}\dot{I}_s &= \dot{V}_s Y_{11} + \dot{V}^*_{if} Y_{12}\\ \dot{I}_{if} &= \dot{V}_{if} Y_{22} + \dot{V}^*_s Y_{21}\end{aligned}\right\} \qquad (4.4)$$

Here,

$$\left.\begin{aligned}Y_{11} &= g_{in,0}\\ Y_{22} &= g_{out,c,0}\\ Y_{12} &= s_{fb,0} = 0.5 g_{fb,k}\\ Y_{21} &= S_c = 0.5 S_k\end{aligned}\right\} \qquad (4.5)$$

An asterisk (*) designates a complex conjugate in the case of an inverting frequency converter.

The above relations solely hold for amplitudes (peak values) and cannot be used to find instantaneous currents and voltages because I_s and V_s apply to processes occurring at one frequency, whereas I_{if} and V_{if} hold for another frequency.

The input current component $\dot{V}^*_{if} Y_{12}$ in Eq. (4.4) reflects the effect of the frequency converter output circuit on this current. In circuits without frequency conversion an increase or a decrease in the load impedance directly leads to a decrease or an increase in the input current. In the case at hand any direct effect is ruled out because the output circuit contains no voltage varying at the input signal frequency. Load reaction shows up as a result of two mutually opposite processes:

— the signal voltage at frequency f_s, after it has been converted, gives rise to an output current at the intermediate frequency. This is what may be called *forward conversion*;

— the i.f. voltage generated in the output circuit acts upon the frequency converter so that frequency conversion takes place, and a current appears in the input circuit at a frequency equal to the signal frequency

$$f_s = kf_{LO} - f_i$$

or

$$f_s = kf_{LO} + f_i$$

This is what may be called *reverse frequency conversion*.

In most frequency converters built around nonreciprocal elements, the feedback conductance is low. This means that in the equivalent circuit of Fig. 4.12 the product $S_{fb,c}V_{if}$ is a very small fraction of $V_s g_{in,0}$. Therefore, one may neglect the effect of reverse conversion and deem the input conductance equal to $g_{in,0}$, that is, to the local-oscillator conductance averaged over a period, g_{in}.

The output conductance of the frequency converter, $g_{out,fc,0}$, is the 'direct' component of the derivative $\partial i_2/\partial v_2$ and is found as the output conductance of the local oscillator, g_{out}, averaged over a period.

Comparison of Fig. 4.12 and Fig. 3.5 brings out a formal identity between the linear models of a frequency converter and of an amplifier. On this basis, it is legitimate to treat frequency converters by invoking the theory of tuned and bandpass amplifiers examined in Chap. 3 on replacing in the respective equations the transconductance of an amplifying elements S (or Y_{21}), by the conversion transconductance S_c (or $Y_{21,c}$), and the output conductance Y_{22} (see Fig. 3.5) by the output conductance of the frequency converter, $g_{out,fc}$ (or $Y_{22,c}$). Notably, for a frequency converter incorporating a bandpass filter the gain is found to be

$$K = S_c mn\rho K_f$$

similar to Eq. (3.95).

As a rule, the conversion transconductance S_c is lower than the transconductance of an amplifier, so the gain of a frequency converter is lower than that of an amplifier without frequency conversion.

Except for the special case taken up a bit later, an increase in the conversion order k causes a decrease in the conversion conductance. This explains why conversion at harmonics is used but seldom. Its use may be warranted, however, if one wishes to lower the local-oscillator frequency. To demonstrate, if

$$f_i = f_s - kf_{LO}$$

then the local-oscillator frequency should be chosen such that

$$f_{LO} = (f_s - f_i)/k$$

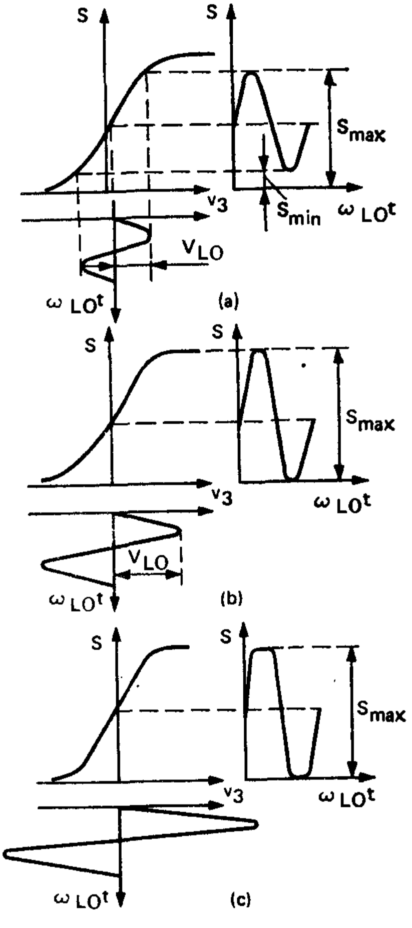


Fig. 4.13

Reduction in the local-oscillator frequency facilitates the onset of self-sustained oscillations in the local oscillator, permits the use of less expensive electron devices and improves frequency stability.

In order to obtain an idea about the relative magnitude of conversion and amplifier transconductances, let us consider Fig. 4.13 which gives an approximate plot of electron-device transconductance as a function of local-oscillator voltage, v_3 (see Fig. 4.11). In the simplest case where the local-oscillator circuit is combined with the input circuit, as in the frequency-converter types shown in Fig. 4.8, the electron-device transconductance, S, is a function of v_1. Consider first-order frequency conversion in which case $S_c = 0.5\,S_1$, where S_1 is the peak value of the variable component of transconductance at the fundamental frequency. It is seen from Fig. 4.13*a* that

$$S_1 \approx 0.5\,(S_{\max} - S_{\min})$$

Therefore

$$S_c \approx 0.25\,(S_{\max} - S_{\min})$$

whereas in an amplifier one may use $S = S_{\max}$. In consequence, $S_c \ll S_{\max}$.

Conversion transconductance can be increased by raising the local-oscillator voltage and by positioning the operating point of the electron device so as to obtain $S_{\min} \approx 0$, as shown in Fig. 4.13*b*. Then, $S_c \approx 0.25\,S_{\max}$, that is, the conversion gain is approximately one-fourth the amplifier gain.

Should the local-oscillator voltage be increased appreciably, as shown in Fig. 4.13*c*, the conversion transconductance will change in about a rectangular manner. From expansion into a Fourier series, it is then seen that

$$S_1 \approx 2\,S_{\max}/\pi \text{ and,}$$

accordingly, $S_c = S_{\max}/\pi$.

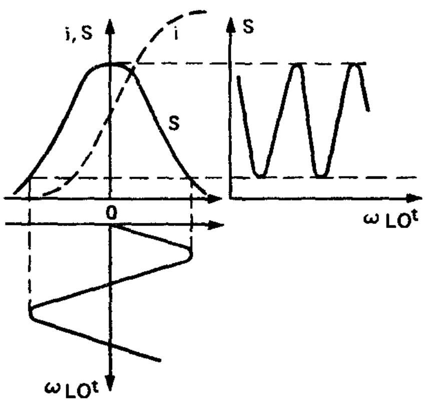


Fig. 4.14

As follows from Sec. 4.3 and Fig. 4.4, first-order conversion is accompanied by higher-order conversion, and this leads to spurious responses. To avoid them, one should choose the operating conditions so that any conversion transconductance harmonics except the fundamental are non-existent, which means that the transconductance should vary as $S = S_0 + S_1 \cos \omega_{\mathrm{LO}} t$.

In consequence, it would be a good plan to use electron devices for which the transconductance-versus-voltage characteristic has a broad nearly linear portion, and to arrange for the local-oscillator voltage to lie within this portion, as shown in Fig. 4.13*a*.

In the case of second-order conversion, the frequency of the desired response is

$$f_s = 2f_{\mathrm{LO}} + f_i$$

or

$$f_s = 2f_{\mathrm{LO}} - f_i$$

(see Fig. 4.4), and the image frequency is

$$f_{\text{im}} = 2f_{\text{LO}} - f_{\text{i}}$$

or

$$f_{\text{im}} = 2f_{\text{LO}} + f_{\text{i}}$$

Also, there will be spurious responses due to first-order conversion, $f_{\text{LO}} - f_{\text{i}}$ and $f_{\text{LO}} + f_{\text{i}}$, and due to third-order conversion, $3f_{\text{LO}} - f_{\text{i}}$ and $3f_{\text{LO}} + f_{\text{i}}$. Responses due to fourth- and higher-order conversion are a far more seldom occurrence.

Second-order conversion transconductance can be enhanced through a proper choice of converter operating conditions. This can, for example, be achieved if the electron device has a volt-ampere characteristic of the form shown by the dashed line in Fig. 4.14. It is seen that the conversion illustrated by the full curve is dominated by the second harmonic, which is an indication that second-order frequency conversion is predominant.

4.8. The Balanced Frequency Converter

Using a differential amplifier as the basis, one can build the frequency converter stage shown in the simplified schematic of Fig. 4.15*a*. In this circuit, the collector voltage is applied to transistors T_1 and T_2 via the centre tap on the inductor of the tuned circuit tuned to the intermediate frequency, f_{i}. The input tuned circuit tuned to the signal frequency is placed between the bases of T_1 and T_2, for which reason the signal voltage is in anti-phase across these two transistors. The local-oscillator voltage at frequency f_{LO} is applied to the base of transistor T_3 and exists in phase at the bases of transistors T_1 and T_2. A decrease or an increase in the T_3 current brings about a corresponding change in the T_1 and T_2 currents and, as a consequence, in their transconductance with the local-oscillator frequency. For this reason, when the signal and local-oscillator voltages are applied simultaneously, frequency conversion takes place.

Since the signal voltage is applied to the bases of T_1 and T_2 in anti-phase, the i.f. current components

$$f_{\text{i}} - f_{\text{LO}} - f_{\text{s}}$$

or

$$f_{\text{i}} = f_{\text{s}} - f_{\text{LO}}$$

are likewise in anti-phase. In the output tuned circuit these currents oppose each other and the i.f. components are therefore combined. The currents at the local-oscillator frequency, which flow through the transistors in phase, cancel each other and produce no voltage in the output circuits.

What we have just examined is referred to as the balanced frequency converter. There also are other modifications of the balanced frequency converter. What is common to all of them is the fact that the signal voltage is applied in phase and the local-oscillator voltage is applied in anti-phase to the two arms of the converter. As is seen in the circuit of Fig. 4.15*b*, the inputs for the signal and local-oscillator voltages are interchanged.

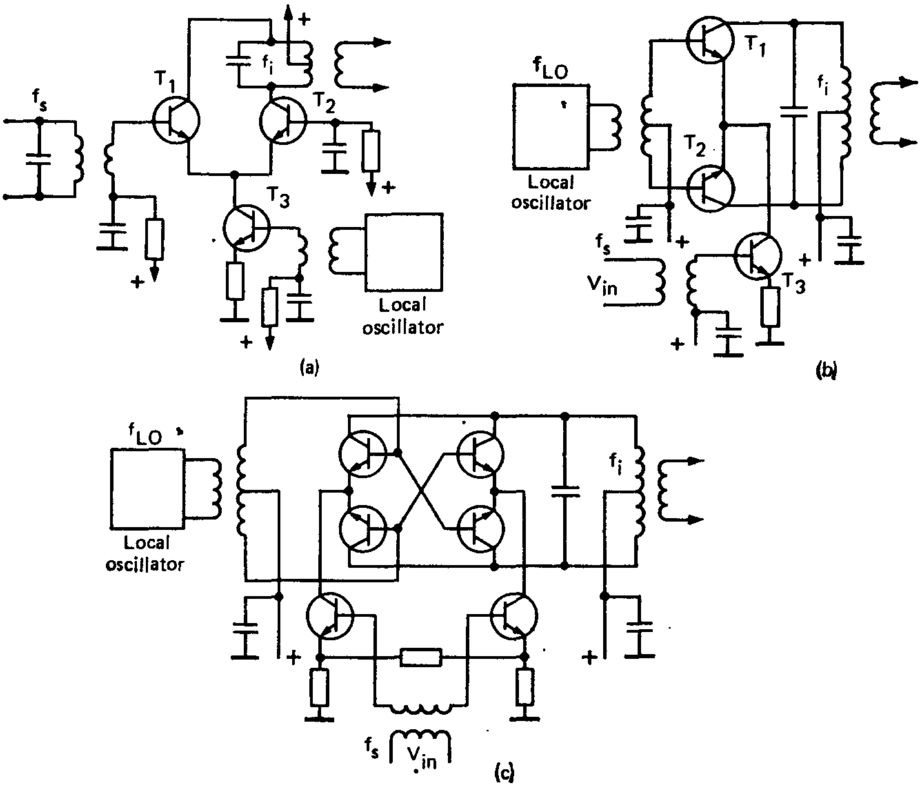


Fig. 4.15

Balanced input circuits can also be arranged for both the signal and the local oscillator. The circuit of a double-balanced frequency converter is shown in Fig. 4.15*c*.

As is known from the theory of amplifiers, an advantage of the balanced or push-pull amplifier is the fact that it suppresses all the even harmonics of the wave being amplified. This advantage is likewise offered by the balanced frequency converter: notably, it cancels out frequencies $2f_{LO} - 2f_{sp}$, where f_{sp} is the frequency of a spurious (interfering) signal. Such a spurious response can arise when a strong interfering signal, $V_{sp} \cos \omega_{sp} t$, happens to find its way to the frequency converter.

In the presence of strong interference, the series in (4.2) should be extended to include second-order components because they can likewise play an important role. In Eq. (4.2), the term not considered before has the form

$$i_m = \frac{1}{2} \frac{\partial^2 f_2 (v_{LO}, E_1, E_2, E_3)}{\partial v_1^2} v_{sp}^2$$

The second derivative of the output current is the first derivative of conversion transconductance, that is, the slope of the curve in Fig. 4.13. Since conversion transconductance varies with local-oscillator frequency, its derivative may likewise be expanded into a Fourier series as follows:

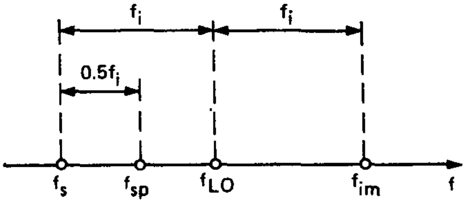


Fig. 4.16

$$\frac{\partial^2 f_2 (v_{LO}, E_1, \ldots)}{\partial v_1^2} = m_0' + \sum_{k=1}^{\infty} m_k \cos k\, \omega_{LO} t$$

Considering v_{sp} and re-arranging, we get

$$i_m = V_{sp}^2 \times 0.5 \Big\{ m_0 (0.5 + 0.5 \cos 2\, \omega_{sp} t) + \sum_{k=1}^{\infty} 0.5 \,[m_k \cos k\omega_{LO} t + 0.5 m_k \cos (k\omega_{LO} + 2\omega_{sp})\, t + 0.5\, m_k \cos (k\omega_{LO} - 2\omega_{sp})\, t] \Big\}$$

The component at angular frequency $2\,(\omega_{LO} - \omega_{sp})$, corresponding to $k = 2$, is especially objectionable when the following two conditions occur at the same time:

— if it falls within the bandwidth of the i.f. amplifier, that is, if $2\,(f_{LO} - f_{sp}) \approx f_i$;

— if an interfering signal at frequency $f_{sp} = f_{LO} - 0.5\, f_i$, producing a component at a frequency close to f_i, finds its way from the antenna to the mixer input.

Figure 4.16 shows the relative position of the signal frequency f_s, the image frequency f_{im}, the local-oscillator frequency f_{LO}, and the spurious (interfering) frequency f_{sp} examined above. The spurious frequency f_{sp} is four times closer to the signal frequency f_s than the image frequency f_{im}, for which reason f_{sp} will be attenuated by the preselector to a lesser extent. In consequence, a spurious response will thus be produced at frequency $f_{sp} \doteq f_{LO} - 0.5\, f_i$.

In the balanced mixer of Fig. 4.15, despite the fact that the interfering signal exists in anti-phase in the two arms (it is $+ v_{sp}$ in one arm and $- v_{sp}$ in the other), the current components i_m are in phase, because $+v_s^2 = -v_{sp}^2$. For this reason, if the mixer is perfectly balanced, the current components i_m will cancel one another in the output turned circuit, and the interfering signal will not be able to find its way into the i.f. amplifier.

4.9. Whistles in Nonlinear Frequency Conversion

When we discussed the theory of frequency conversion in Sec. 4.1, it was assumed that the current in a circuit with periodically varying parameters was related to the signal voltage in a linear manner. Under this assumption, with the signal frequency being f_s, the current spectra contain components at frequencies $| kf_{LO} \pm f_s |$. However, it is not always that the incoming signals are so weak that the converter (mixer) nonlinearity with respect to them may be neglected. Notably, in broadcast reception the signals of local radio stations may be so strong that one is forced to consider nonlinear effects.

On extending the series in (4.2) which represented the converter current in general form in Sec. 4.7, note the terms of the form

$$\frac{1}{m!} \frac{\partial m (v_{LO}, E_1, \ldots)}{\partial v_1^m} v_s^m \tag{4.6}$$

Consideration of the second-order term ($m = 2$) brings out the existence of a spurious reception channel. Reasoning along a similar line, note that the derivative of any order, $\partial^n f (v_{LO}, \ldots)/\partial v_1^n$, much as the derivatives $\partial f (v_{LO}, \ldots)/\partial v_1$ and $\partial^2 f (v_{LO}, \ldots)/\partial v_1^2$, varies in the general case nonlinearly with the local-oscillator frequency and may be expanded into a Fourier series. Note also that when

$$v_s = V_s \cos \omega_s t$$

there appears, in Eq. (4.6), the factor $\cos^m \omega_s t$ which can, in turn, be expanded into a trigonometric series and will then contain a term proportional to $\cos m\omega_s t$. Thus, according to (4.6), the converter current will contain components proportional to the product $\cos k\omega_{LO} t \times \cos m\omega_s t$. For this reason, the current spectrum will contain components at frequencies $f_i' = kf_{LO} - mf_s$ or $f_i' = mf_s - kf_{LO}$. If f_1' is close to f_i, that is, if $| kf_{LO} - mf_s | \approx f_i$, then two conversion products of the same signal will fall within the bandwidth of the i.f. amplifier: one as a result of first-order conversion at f_i and the other (a weaker one as all nonlinear high-order conversion .products) at frequency f_i'.

Owing to the fact that the principal signal at f_i has a higher level in the i.f. section than the signal at f_1', it will be received, but the signal at the shifted frequency f_i', present in its spectrum, will appear as interference.

When the two signals beat together, the amplitude and phase of the resultant wave varies at a frequency which is the difference between the frequencies of the two signals. What we thus have is parasitic modulation of the valid signal by the beat frequency $f_b = f_i - f_i'$. After detection (demodulation), this modulation will be perceived in aural reception as a whistle or tweet at frequency f_b. With other types of signals reproduced in a different manner, the disturbance may take a different form, this phenomenon characteristic of superheterodyne reception as a whole is referred as *whistles* or *tweets* all the same.

Whistles will occur if f_i' falls within the passband of the amplifier tuned to f_i, that is, if the difference frequency $|f_i - f_i'|$ does not exceed half the amplifier bandwidth, $B/2$. If $B/2 \ll f_i$, the interference will arise at

$$f_i \approx f_i' = kf_{LO} - mf_s \qquad (4.7)$$

If $f_i = f_s - f_{LO}$ or $f_i = f_{LO} - f_s$, that is, if $f_{LO} = f_s - f_i$ or $f_{LO} = f_s + f_i$, then on substituting in Eq. (4.7), we obtain

$$f_s \approx f_i\,(k+1)/(k-m)$$

or

$$f_s \approx f_i\,(k-1)/(m-k)$$

If the frequency f_s thus found falls within the frequency range of a receiver, then whistles may be produced when the receiver is tuned to that frequency.

4.10. The Diode Frequency Converter

A diode may be used for frequency conversion similarly to any other electron device possessing a nonlinear characteristic. Owing to its low noise level, ability to operate at the highest radio frequencies, and simple construction, the diode frequency converter is used in nearly all SHF receivers.

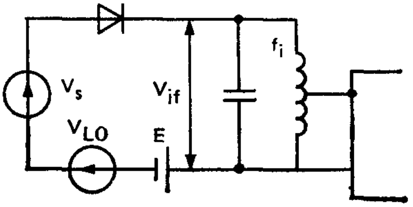


Fig. 4.17

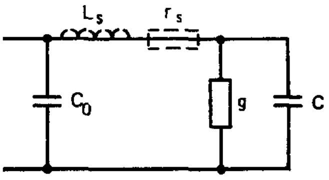


Fig. 4.18

The circuit of a diode frequency converter is shown in Fig. 4.17, and an equivalent circuit of the diode, in Fig. 4.18. Here g and C

are the conductance and capacitance of the *n-p* junction; r_s and L_s are the spurious (lead) resistance and inductance; and C_0 is the capacitance of the diode holder. In the UHF and SHF bands, the L_s and r_s of state-of-the-art mixer diodes are very low, so that they produce almost no effect and may therefore be neglected.

The bias voltage source E (see Fig. 4.17) serves to position the operating point of the diode.

In approximate form, plots relating the diode conduction current i, differential conductance g, capacitance C (considering C_0 as well), and charge q as a function of the applied voltage are shown in Fig. 4.19.

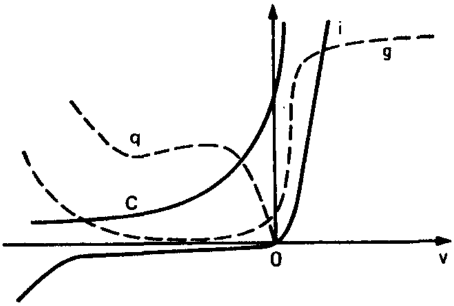


Fig. 4.19

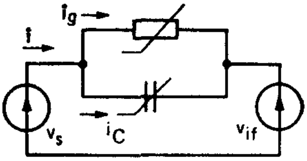


Fig. 4.20

As we did in Sec. 4.7, let us deem, in analysing the diode converter, that v_s and v_{if} are small in comparison with v_{LO}. This assumption is based on the following:

1. The local-oscillator voltage should have an amplitude sufficiently large for changes in the current to encompass the nonlinear portion of the diode characteristic; this is a necessary condition for frequency conversion to take place.

2. Amplification is effected mainly past the frequency converter, that is, in the i.f. amplifier; therefore, the input and output converter voltages are low.

When the signal voltage v_s and the i.f. output voltage v_{if} are relatively low, the diode does not practically display its nonlinearity, and acts as a variable-parameter linear network with regard to the signal. Accordingly, the equivalent circuit of a diode frequency converter may be drawn as shown in Fig. 4.20.

Receiver selectivity is mainly provided past the frequency converter, that is, by the i.f. amplifier. The circuits ahead and up to the frequency converter mainly attenuate spurious responses, whereas other interfering frequencies, including those markedly exceeding in level the wanted signal, may reach the frequency converter.

The interaction of the local-oscillator signal, the wanted signal, and interfering signals in nonlinear circuits may give rise not only to the intermodulation products already examined, but to more complex components whose frequencies may likewise be close to the intermediate frequency. Falling within the bandwidth of the i.f. amplifier, they show up as unavoidable spurious responses, thus imparing receiver selectivity. It is therefore important that the frequency converter be linear with regard to spurious responses, and this depends on the properties and operating conditions of the diode.

Linearity presumes that the diode parameters *g* and *C* (see Fig. 4.19) are independent of the voltages being converted. In other words, variations in these parameters should solely be decided by the local-oscillator voltage which is by several orders of magnitude higher than the signal voltage and the interfering voltages. In view of this, the circuit seen by the local-oscillator (see Fig. 4.17) may be regarded as a nonlinear load, similarly to a frequency converter built around a nonreciprocal electron device (see Sec. 4.7). Then, in order to analyse the diode frequency converter, we should determine the manner in which *g* and *C* will vary in response to the action of the local oscillator, as well as the input resistance of the diode seen by the local-oscillator and the power that the mixer draws from the local oscillator.

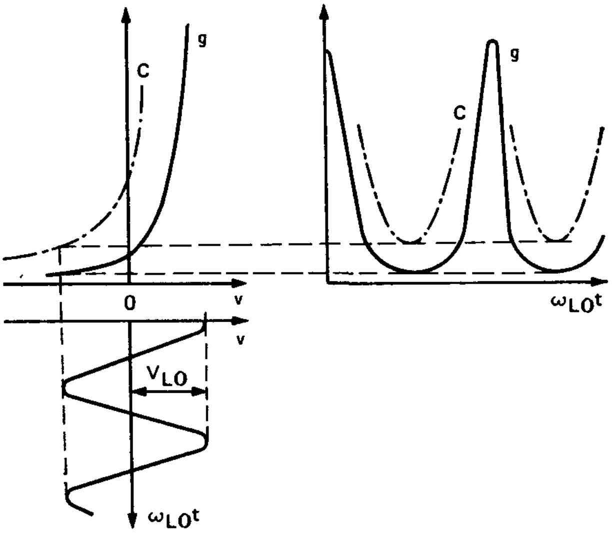


Fig. 4.21

A plot of variations in the diode conductance *g* and the diode capacitance *C* under the action of the local-oscillator voltage $v_{LO}=$

$V_{\mathrm{LO}} \cos \omega_{\mathrm{LO}} t$ is shown in Fig. 4.21. These variations may be expanded into Fourier series as follows:

$$g(t) = g_0 + \sum_{k=1}^{\infty} g_k \cos k\omega_{\mathrm{LO}} t$$

$$C(t) = C_0 + \sum_{k=1}^{\infty} C_k \cos k\omega_{\mathrm{LO}} t$$

On denoting the a.c. voltage in the diode circuit in Fig. 4.20 as v, the current in the diode circuit may be defined as

$$i = vg + C\,dv/dt + v dC/dt$$

Let

$$v_s = V_s \cos(\omega_s t + \varphi_s)$$

and

$$v_{if} = V_{if} \cos(\omega_i t + \varphi_{if})$$

The phase angle φ_{if} depends on the relative magnitude of the capacitive and resistive components of the diode conductance and the phase angle of load conductance (in Fig. 4.17, the load is a tuned circuit). Let us substitute for g and C and also $v = v_s + v_{if}$ in the expression for i, replace the products of trigonometric functions by functions of sum and difference phase angles, and collect the terms. As a result, we will obtain

$$\begin{aligned}
i = {} & V_s [g_0 \cos(\omega_s t + \varphi_s) - \omega_s C_0 \sin(\omega_s t + \varphi_s)] \\
& + V_{if} [g_0 \cos(\omega_i t + \varphi_{if}) - \omega_i C_0 \sin(\omega_i t + \varphi_{if})] \\
& + V_s \sum_{k=1}^{\infty} 0.5 g_k \{\cos(k\omega_{\mathrm{LO}} + \omega_s) t + \varphi_s] \\
& + \cos[k\omega_{\mathrm{LO}} t - \omega_s) t - \varphi_s]\} + V_{if} \sum_{k=1}^{\infty} 0.5 g_k \{\cos \\
& \times [(k\omega_{\mathrm{LO}} + \omega_i) t + \varphi_{if}] + \cos[(k\omega_{\mathrm{LO}} - \omega_i) t - \varphi_{if}]\} \\
& - V_s \sum_{k=1}^{\infty} 0.5 C_k \{(k\omega_{\mathrm{LO}} + \omega_s) \sin[(k\omega_{\mathrm{LO}} + \omega_s) t + \varphi_s] \\
& + (k\omega_{\mathrm{LO}} - \omega_s) \sin[(k\omega_{\mathrm{LO}} - \omega_s) t - \varphi_s]\} \\
& - V_{if} \sum_{k=1}^{\infty} 0.5 C_k \{k(\omega_{\mathrm{LO}} + \omega_i) \sin[(k\omega_{\mathrm{LO}} + \omega_i) t + \varphi_{if}] \\
& + (k\omega_{\mathrm{LO}} - \omega_i) \sin[(k\omega_{\mathrm{LO}} - \omega_i) t - \varphi_{if}]\} \qquad (4.8)
\end{aligned}$$

For a noninverting frequency converter (see Sec. 4.1),

$$\omega_i = \omega_s - k\omega_{LO} \quad \text{or} \quad \omega_i = \omega_s + k\omega_{LO}$$

Accordingly,

$$\omega_s = \omega_i \pm k\omega_{LO}$$

From Eq. (4.8), the components at these frequencies are

$$\begin{aligned} i_{if} = {} & V_s\,[0.5\, g_k \cos(\omega_i t + \varphi_s) - 0.5\, \omega_i C_k \sin(\omega_i t \\ & + \varphi_s)] + V_{if}\,[g_0 \cos(\omega_i t + \varphi_{if}) \\ & - \omega_i C_0 \sin(\omega_i t + \varphi_{if})] \\ i_s = {} & V_s\,[g_0 \cos(\omega_s t + \varphi_s) - \omega_s C_0 \sin(\omega_s t + \varphi_s)] \\ & + V_{if}\,[0.5 g_k \cos(\omega_s t + \varphi_{if}) - 0.5\, \omega_s C_k \sin(\omega_s t + \varphi_{if})] \end{aligned}$$

Proceeding from the above expressions, the complex current amplitudes are found to be

$$\begin{aligned} \dot{I}_{if} &= V_s\,(0.5\, g_k + j\, 0.5\, \omega_i C_k) + V_{if}\,(g_0 + j\omega_i C_0) \\ \dot{I}_s &= V_s\,(g_0 + j\omega_s C_0) + V_{if}\,(0.5\, g_K + j\, 0.5\, \omega_s C_k) \end{aligned} \tag{4.9}$$

For an inverting frequency converter,

$$\omega_i = k\omega_{LO} - \omega_s$$

and

$$\omega_s = k\omega_{LO} - \omega_i$$

Hence, it follows from Eq. (4.8) that

$$\begin{aligned} i_{if} = {} & V_s\,[0.5\, g_k \cos(\omega_i t - \varphi_s) - 0.5\, \omega_i C_k \sin(\omega_i t - \varphi_s)] \\ & + V_{if}\,[g_0 \cos(\omega_i t + \varphi_{if}) - \omega_i C_0 \sin(\omega_i t + \varphi_{if})] \\ i_s = {} & V_s\,[g_0 \cos(\omega_s t + \varphi_s) - \omega_s C_0 \sin(\varphi_s t + \varphi_s)] \\ & + V_{if}\,[0.5\, g_k \cos(\omega_s t - \varphi_{if}) - 0.5\, \omega_s C_k \sin(\omega_s t - \varphi_{if})] \end{aligned}$$

As is seen from the foregoing, the current components have phase angles which take signs opposite to those of the phase angles of the input voltages, and this implies that the respective complex amplitudes have an imaginary part with a sign opposite to its sign in the case of a noninverting frequency converter. In other words, these components have complex conjugate amplitudes, $\dot{V}_s^*$ and $\dot{V}_{if}^*$. Thus, in contrast to Eq. (4.9), the formulas for the complex amplitudes take the form

$$\begin{aligned} \dot{I}_{if} &= \dot{V}_s^*\,(0.5\, g_k + j\, 0.5\, \omega_i C_k) + \dot{V}_{if}\,(g_0 + j\omega_i C_0) \\ \dot{I}_s &= \dot{V}_s\,(g_0 + j\omega_s C_0) + \dot{V}_{if}^*\,(0.5\, g_k + j\, 0.5\, \omega_s C_k) \end{aligned} \tag{4.10}$$

Let $0.5\,C_k = C_c$ be the conversion capacitance, and $0.5\,g_k = g_c$ be the conversion conductance. Also, let us adopt the following notation.

$$\begin{aligned} \dot{Y}_{11} &= g_0 + j\omega_s C_0 \\ \dot{Y}_{12} &= g_c + j\omega_s C_c \\ \dot{Y}_{21} &= g_c + j\omega_i C_c \\ \dot{Y}_{22} &= g_0 + j\omega_i C_0 \end{aligned} \tag{4.11}$$

Then Eqs. (4.9) and (4.10) will, respectively, take the following form

$$\left.\begin{aligned} \dot{I}_{\mathrm{if}} &= \dot{V}_s \dot{V}_{21} + \dot{V}_{\mathrm{if}} \dot{Y}_{22} \\ \dot{I}_s &= \dot{V}_s \dot{Y}_{11} + \dot{V}_{\mathrm{if}} \dot{Y}_{12} \end{aligned}\right\} \tag{4.9a}$$

and

$$\left.\begin{aligned} \dot{I}_{\mathrm{if}} &= \dot{V}_s^* \dot{Y}_{21} + \dot{V}_{\mathrm{if}} \dot{Y}_{22} \\ \dot{I}_s &= \dot{V}_s \dot{Y}_{11} + \dot{V}_{\mathrm{if}}^* \dot{Y}_{12} \end{aligned}\right\} \tag{4.10a}$$

Proceeding from the above expressions, one may design a frequency converter by using, as in Sec. 4.7, an equivalent circuit in the form of a nonlinear two-port. The parameter $\dot{Y}_{21}$ takes care of the conversion of the signal current to the i.f. current, and the parameter $\dot{Y}_{12}$ reflects the effect of the load on the input current in the case of forward and reverse frequency conversion.

The first line in Eqs. (4.9a) offers a means for finding the voltage gain of a noninverting frequency converter. Defining $\dot{V}_{\mathrm{if}}$ as the voltage drop across the load impedance, $\dot{Z}_L$, at the converter output and noting its sign relative to the input emf source, that is, as $\dot{V}_{\mathrm{if}} = -\dot{I}_{\mathrm{if}} \dot{Z}_L$, we get

$$\dot{I}_{\mathrm{if}} = \dot{V}_s \dot{Y}_{21} - \dot{I}_{\mathrm{if}} \dot{Z}_L \dot{Y}_{22}$$

Hence

$$\dot{I}_{\mathrm{if}} = \dot{V}_s \dot{Y}_{21} / (1 + \dot{Z}_L \dot{Y}_{22})$$

Accordingly, the output voltage of the frequency converter is

$$\dot{V}_{\mathrm{if}} = -\dot{I}_{\mathrm{if}} \dot{Z}_L = -\dot{V}_s \dot{Y}_{21} \dot{Z}_L / (1 + \dot{Z}_L \dot{Y}_{22}) \tag{4.12}$$

For an inverting frequency converter, $\dot{V}_s$ in the above equation should be replaced by $\dot{V}^*_s$; this will not affect the output voltage and is solely related to its phase shift from the signal voltage.

The gain of the frequency converter is

$$\dot{K}_{\mathrm{C}} = \dot{Y}_{21}\dot{Z}_{\mathrm{L}}/(1 + \dot{Z}_{\mathrm{L}}\dot{Y}_{22})$$

or, in a different way

$$\dot{K}_{\mathrm{C}} = \dot{Y}_{21}/(\dot{Y}_{\mathrm{L}} + \dot{Y}_{22}) \tag{4.13}$$

where

$$\dot{Y}_{\mathrm{L}} = 1/\dot{Z}_{\mathrm{L}}$$

For a noninverting frequency converter Eq. (4.12) may be written as

$$\dot{V}_{\mathrm{C}} = -\dot{V}_{\mathrm{S}}\dot{K}_{\mathrm{C}} \tag{4.14}$$

and for an inverting converter

$$\dot{V}_{\mathrm{C}} = -\dot{V}_{\mathrm{S}}^{*}\dot{K}_{\mathrm{C}} \tag{4.15}$$

In view of Eq. (4.14), we obtain from Eq. (4.9*a*) that

$$\dot{I}_{\mathrm{S}} = \dot{V}_{\mathrm{S}}\dot{Y}_{11} - \dot{V}_{\mathrm{S}}\dot{K}_{\mathrm{C}}\dot{Y}_{12}$$

Hence, the input admittance of a noninverting frequency converter

$$\dot{Y}_{\mathrm{in}} = \dot{I}_{\mathrm{S}}/\dot{V}_{\mathrm{S}} = \dot{Y}_{11} - \dot{Y}_{12}\dot{K}_{\mathrm{C}} \tag{4.16}$$

On the basis of Eq. (4.15), for an inverting frequency converter we have

$$\dot{V}_{\mathrm{C}}^{*} = -\dot{V}_{\mathrm{S}}\dot{K}_{\mathrm{C}}^{*}$$

In consequence,

$$\dot{I}_{\mathrm{S}} = \dot{V}_{\mathrm{S}}\dot{Y}_{11} - \dot{V}_{\mathrm{S}}\dot{Y}_{12}\dot{K}_{\mathrm{C}}^{*}$$

Hence,

$$\dot{Y}_{\mathrm{in}} = \dot{Y}_{11} - \dot{Y}_{12}\dot{K}_{\mathrm{C}}^{*} \tag{4.17}$$

The diode of a frequency converter is most often used under any one of the following sets of operating conditions.

1. The local-oscillator voltage varies predominantly within the forward-current region and moves into the reverse-current region for only a part of a cycle, in which case use is made of a diode having a low capacitance. In a case like that, the primary role is played by the nonlinear resistance of the diode whereas its capacitance only slightly affects the process of conversion. Quite aptly, this is what is known as a resistive frequency converter.

2. Owing to the fact that a negative bias voltage (*E* in Fig. 4.17) is applied to the diode, the local-oscillator voltage mainly varies

within the negative region, and use is made of a diode with a relatively high nonlinear capacitance, that is, a varactor. The diode resistance plays a minor role. This is a capacitive frequency converter.

4.11. The Resistive Diode Frequency Converter

Diode frequency converters (often called diode mixers) are widely used in receivers, especially at microwave frequencies. An equivalent circuit of such a converter, along with its input tuned circuit, equivalent input signal source, output tuned circuit, and the input conductance of the succeeding i.f. amplifier is shown in Fig. 4.22. Let us apply the theory set forth above to this circuit. Neglecting the capacitances, we have from Eq. (4.11)

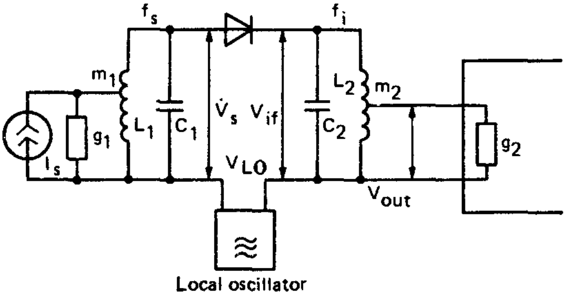


Fig. 4.22

$$Y_{11} = Y_{22} = g_0$$

$$Y_{12} = Y_{21} = g_c$$

From Eqs. (4.13) and (4.16), we obtain

$$\dot{K}_c = g_c/(\dot{Y}_L + g_0)$$

$$\dot{Y}_{in} = g_0 - g_c^2\,(g_0 + \dot{Y}_L)$$

Then,

$$\dot{Y}_L = g_{ckt_2} + m_2^2 g_2 + jb_2$$

where g_{ckt_2} is the equivalent loss conductance of the output tuned circuit and

$$b_2 = \omega_i C_2^{\cdot} - 1/\omega_i L_2$$

or, in a different way,

$$b_2 = \rho_2 y_2$$

where

$$y_2 = f_1/f_2 - f_2/f_1$$
$$\rho_2 = (L_2/C_2)^{1/2}$$
$$f_2 = 1/2\,\pi\,(L_2C_2)^{1/2}$$

is the resonance frequency of the output tuned circuit.

As follows from Fig. 4.22, the input signal voltage of the frequency converter is

$$\dot{V}_s = \dot{I}_s m_1/(g_{\text{ckt}1} + m_1^2 g_1 + jb_1 + \dot{Y}_{\text{in}}) \tag{4.18}$$

where

$g_{\text{ckt}1}$ = loss conductance of the input tuned circuit

$$b_1 = \omega_s C_1 - 1/\omega_s L$$

or, in a different way,

$$b_1 = \rho_1 y_1$$

where

$$y_1 = f_s/f_1 - f_1/f_s$$
$$\rho_1 = (L_1C_1)^{1/2}$$
$$f_1 = 1/2\,\pi\,(L_1C_1)^{1/2}$$

As follows from Norton's theorem, in Eq. (4.18)

$$I_s = E_s g_1$$

where E_s is the signal source emf.

The output voltage of the converter can be found as

$$\dot{V}_{\text{out}} = \dot{V}_s \dot{K}_c m_2$$

On defining the overall gain as

$$K = \dot{V}_{\text{out}}/E_s$$

we obtan

$$\dot{K} = \frac{g_1 g_c m_1 m_2}{(g_0 + g_{\text{ckt}1} + m_1^2 g_1 + jb_1)\,(g_0 + g_{\text{ckt}2} + m_2^2 g_2 + jb_2) - g_c^2} \tag{4.19}$$

Equation (4.19) can be used to calculate the amplitude and phase responses of the converter. At resonance ($b_1 = 0$ and $b_2 = 0$), we have

$$K_0 = \frac{g_1 g_c m_1 m_2}{(g_0 + g_{\text{ckt}\,1} + m_1^2 g_1)\,(g_0 + g_{\text{ckt}\,2} + m_2^2 g_2) - g_c^2} \tag{4.20}$$

It can be seen from Eq. (4.20) that m_1 and m_2 may be chosen so as to maximize K_0. Let us find a maximum K_0 for a simplified case where the input and output tuned-circuit losses represented by

$g_{\text{ckt}1}$ and $g_{\text{ckt}2}$ are relatively small and may be neglected, which is often the case in practice. Then,

$$K_0 \approx \frac{g_2 g_c m_1 m_2}{(g_0 + m_1^2 g_1)(g_0 + m_2^2 g_2) - g_c^2} \tag{4.21}$$

On designating $m_1 g_1^{1/2} = x_1$ and $m_2 g_2^{1/2} = x_2$, let us rewrite Eq. (4.21) as

$$K_0 = g_c (g_1/g_2)^{1/2} \frac{x_1 x_2}{(g_0 + x_1^2)(g_0 + x_2^2) - g_c^2}$$

Since K_0 is equally dependent on x_1 and x_2, a maximum value of K_0 in terms of these variables may occur when $x_1 = x_2 = x$. Therefore, in order to find the maximum, it is legitimate to write K_0 as

$$K_0 = g_c (g_1/g_2)^{1/2} \frac{x_2}{(g_0 + x^2)^2 - g_c^2}$$

On equating to zero the derivative of K_0 with respect to x^2, we find

$$x_{\text{opt}} = (g_0^2 - g_c^2)^{1/4}$$

Therefore,

$$m_{1,\text{opt}} = x_1 g_1^{1/2} = [(g_0^2 - g_c^2/g^2]^{1/4}$$

and

$$m_{2,\text{opt}} = x_2/g_2^{1/2} = [(g_0^2 - g_c^2)/g_2^2]^{1/4}$$

Given optimal values of m_1 and m_2,

$$K_{0,\,\max} = \frac{1}{2} (g_1/g_2)^{1/2} \frac{g_c}{g_0 + (g_0^2 - g_c^2)^{1/2}}$$

On denoting $g_c/g_0 = \mu_c$, the above equation may be re-cast as

$$K_{0,\,\max} = \frac{1}{2} (g_1/g_2)^{1/2} \frac{\mu_c}{1 + (1 - \mu_c^2)^{1/2}} \tag{4.22}$$

With μ_c tending to unity, $K_{0,\max}$ tends to 0.5 $(g_1/g_2)^{1/2}$, which is another way of saying that the diode parameters do not affect the gain. This is the case of an ideal, lossless frequency converter. The gain obtained above corresponds to an optimal amount of coupling from the signal source of conductance g_1 to the next circuit of conductance g_2 via an ideal (lossless) transformer. The actual gain is always lower than the one obtained above because for real diodes $g_c < g_0$ and, accordingly, $\mu_c < 1$, and also because unaccounted losses occur in resonant circuits.

As an example, consider a frequency converter which uses a diode for which the volt-ampere characteristic is shown by the full line in the plot of Fig. 4.23. In this example, the forward and reverse currents are, to a first approximation, represented by straight lines. Also, the forward conductance g_a and the reverse conductance g_r

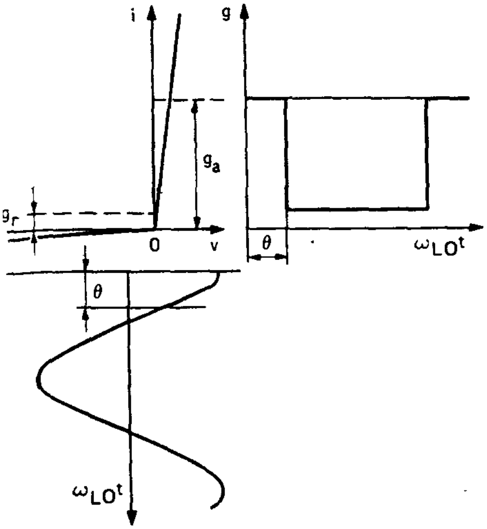


Fig. 4.23

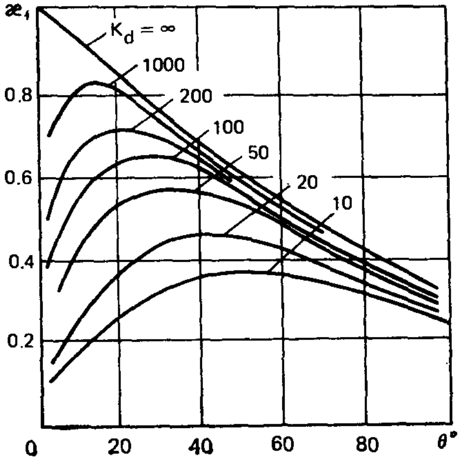


Fig. 4.24

are constant, as shown by the dashed lines. The angle θ represents the time interval during which the diode is conducting. Noting that the function being analysed is an even one, we obtain for first-order conversion

$$g_0 = (1/\pi)\left(\int_0^{\theta} g_a\, d\omega t + \int_{\theta}^{\pi} g_r\, d\omega t\right)(1/\pi)\,[g_a\theta + g_r(\pi - \theta)]$$

$$g_c = 0.5 g_1 = 0.5\,(2/\pi)\left(\int_0^{\theta} g_a \cos\omega_{LO}t\, d\omega_{LO}t\right.$$

$$\left. + \int_{\theta}^{\pi} g_r \cos\omega_{LO}t\, d\omega_{LO}t\right) = (1/\pi)\sin\theta\,(g_a - g_r)$$

Therefore,

$$\mu_c = (\sin\theta/\theta)\,[1 + (\pi/\theta)\, g_r/(g_a - g_r)]^{-1}$$

The ratio g_a/g_r characterizes the rectifying properties of the diode and its quality. Let us denote g_a/g_r as K_D. Then

$$\mu_c = (\sin\theta/\theta)\,\frac{\theta/\pi}{(\theta/\pi) + 1\,(K_D - 1)} \tag{4.23}$$

Let us re-write Eq. (4.22) as

$$K_{0,\mathrm{max}} = 0.5\,(g_1/g_2)^{1/2}\varkappa_1$$

where

$$\varkappa_1 = \mu_c/[(1 + (1 - \mu_c^2)^{1/2}] \qquad (4.24)$$

Here, $\varkappa_1$ is the efficiency of a diode mixer. A plot of $\varkappa_1$ as a function of θ for several values of K_D, as computed by Eqs. (4.23) and (4.24), is shown in Fig. 4.24.

It is seen from the above equations and Fig. 4.24 that $\varkappa_1$ is always less than unity which is another way of saying that ideal conversion is never obtained. Also, it follows from Fig. 4.24 that the diode can be operated in an optimal manner such that $\varkappa_1$ and, as a consequence, K_0 are maximal. The angle of current flow (the operating or conduction angle) of the diode, θ, can be adjusted by varying the bias voltage E (see Fig. 4.17) and the amplitude of the local-oscillator voltage, V_{LO}.

4.12. Sources of Noise in a Frequency Converter

The frequency converter is one of the early stages in a receiver for which reason its noise has a direct bearing on receiver sensitivity. The sources of noise in a frequency converter are the same as they are in the other stages. They are dealt with in a course on electron devices and have been examined in Chap. 3 of this text.

In a frequency converter built around a nonreciprocal electron device (see Secs. 4.6 and 4.7), noise arises due to fluctuations of charge-carrier streams in the electron device itself and due to thermal fluctuations in the associated circuits.

Of the broad noise spectrum originating in the electron device of a frequency converter, only that part reaches the input of the i.f. section which falls within the bandwidth of the i.f. amplifier. As the theory of electron devices tells us, the mean square of noise current is proportional to the current in the electron-device circuit. It is a special feature of a frequency converter that this current varies at the local-oscillator frequency, so the mean square of the noise spectrum portion falling within the bandwidth is proportional to the average (direct) current.

Noise current in the input circuit of a frequency converter has a wider spectrum than at the input to the i.f. amplifier; its width is decided by the passband of the r.f. section ahead of the frequency converter. Noise can be calculated, using a circuit similar to that shown in Fig. 3.21. If its selectance or discrimination against the spurious response frequencies is not high enough, the spectrum components having the same frequencies as the spurious channels past the frequency converter will fall within the bandwidth of the i.f. section, and the overall noise level at the frequency converter output will rise.

If the input conductance of the frequency converter in Fig. 4.11 is g_{in} and the input voltage is V_1, then the power drawn from the signal source will be

$$P_{\mathrm{in}} = V_1^2 g_{\mathrm{in}}$$

If the output voltage is V_2 and the load conductance is g_2, then

$$P_{\mathrm{out}} = V_2^2 g_2$$

Therefore, in view of the fact that

$$V_2/V_1 = K_{\mathrm{c}}$$

the power gain of the frequency converter may be written as

$$K_P = P_{\mathrm{out}}/P_{\mathrm{in}} = K_{\mathrm{c}}^2\,(g_2/g_1)$$

Let the mean square of the converter output noise voltage be $\overline{V_{\mathrm{n,m}}^2}$. Then the converter noise output power will be

$$P_{\mathrm{n,m}} = \overline{V_{n,\mathrm{c}}^2}\, g_2$$

If the noise power coming from the signal source is $P_{\mathrm{n,0}}$, then the total noise output power will be

$$P_{\mathrm{n,out}} = P_{\mathrm{n,0}} K_P + \overline{V_{\mathrm{n,c}}^2}\, g_2$$

Hence, the noise figure of the frequency converter is

$$N_{\mathrm{c}} = P_{\mathrm{n,out}}/P_{\mathrm{n,0}} K_P = 1 + (\overline{V_{\mathrm{n,c}}^2}\, g_2/P_{\mathrm{n,0}} K_{\mathrm{P}})$$

or, in view of the expression for K_P,

$$N_{\mathrm{c}} = 1 + [(\overline{V_{\mathrm{n,c}}^2}\, g_1/P_{\mathrm{n,0}} K_{\mathrm{C}}^2)]$$

If, instead of a frequency converter, the same electron device were used in an amplifier of gain K, then the noise figure would be

$$N = 1 + (\overline{V_{\mathrm{n}}^2}\, g_1/P_{\mathrm{n,0}} K^2)$$

where $\overline{V_{\mathrm{n}}^2}$ is the mean square of amplifier noise voltage and K is the amplifier gain. Where $\overline{V_{\mathrm{n}}^2}$ and $\overline{V_{\mathrm{n,c}}^2}$ are close in value, the difference in the noise figures is due to the difference between K_{c} and K. It has been shown in Sec. 4.7 that the conversion gain, with all of the capabilities of the electron device utilized to the utmost, is about one-fourth of the amplifier gain. It follows then that the noise figure of a frequency converter is substantially greater than that of an amplifier.

The noise figure may also go up because at the frequency f_{i} of the output tuned circuit the usual noise originating in the electron device is augmented by the noise falling within the passband due to frequency conversion, thus causing a very marked increase in the noise figure of the receiver. This can be avoided if the frequency converter is preceded by a low-noise, high-gain amplifier, as follows from Eq. (1.26).

The noise of a diode frequency converter consists of the following components.

1. Input-circuit thermal noise translated by conversion from the incoming signal frequency range to the passband of the i.f. amplifier. The mean square of this noise current is, according to Eq. (1.5), given by $4\,kTBg_{\text{ckt}}$, where g_{ckt} is the loss conductance of the input tuned circuit. This current flows in a circuit composed of three shunt conductances, namely the tuned-circuit conductance g_1 (see Fig. 4.22), referred signal-source conductance $m_1^2 g_1$, and the converter input conductance g_{in}. The latter is associated with the passage of input-circuit thermal noise through the converter. As a result of forward conversion, this noise produces an i.f. voltage at the output, whereas as a result of reverse conversion it is again translated into the frequency range of the input signal.

2. Image-frequency input-circuit thermal noise translated to the passband of the i.f. amplifier. This noise plays a prominent role when the input circuit has a broad bandwidth enclosing the image channel; it is likewise associated with reverse frequency conversion.

3. Converter output-circuit thermal noise within the passband of the i.f. amplifier. The mean square of this noise current is $4kTB(g_{\text{ckt}2} + m_2^2 g_2)$, where $g_{\text{ckt}2}$ is the conductance of the output tuned circuit, and $m_2^2 g_2$ is the transferred load conductance. If the load circuit contains further sources of noise, they, too, must be taken into consideration. This noise current flows in a circuit composed of $g_{\text{ckt}2}$, $m_2^2 g_2$ and the converter conductance on the output-circuit side. The latter is defined similarly to the input conductance on the signal-source side and takes care of the passage of output-circuit noise; owing to reverse conversion this noise is translated back into the signal frequency range and produces across the input tuned circuit a voltage which is then translated by forward conversion to the passband of the i.f. amplifier. Some role may also be played by the translation of output-circuit noise to the image channel. If this is the case, a noise voltage is then generated in the input circuit to be again converted to i.f. noise. It is seen, therefore, that the input circuit may affect the resultant output-circuit thermal noise voltage in a complex manner.

4. The diode noise current which produces in the output circuit a voltage falling within the passband of the i.f. amplifier. This current contains a thermal component, but it is mainly determined by shot noise for which (as is known from the theory of electron devices) the mean square is $2eI_0B$, where e is the charge on an electron, and I_0 is the direct component of diode current.

5. The diode noise current producing in the input circuit voltages falling within the signal frequency range and in the image channel. By forward conversion, these voltages are translated to the passband of the i.f. amplifier.

The converter noise figure N_{c} is determined, as has been done earlier, on the assumption that the signal source noise output power reaching the converter input is $P_{\mathrm{n},0}$. The noise figure is defined

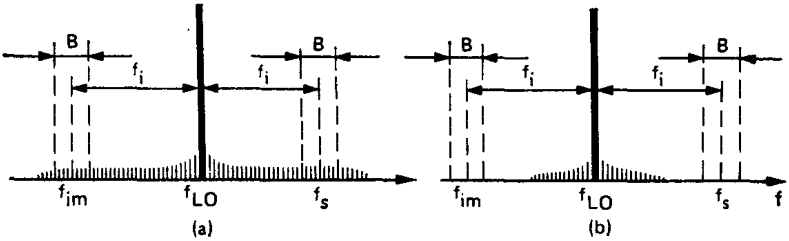


Fig. 4.25

as the ratio of the sum of the powers (or of the mean squares of voltages) due to the factors listed above, to the power associated with the signal source, $P_{\mathrm{n},0}$.

Assuming a perfect match between the coupling circuits and according to Eq. (1.26), the noise figure for both the frequency converter and the succeeding i.f. amplifier is

$$N = N_{\mathrm{c}} + (N_{\mathrm{amp}} - 1)/K_{P,\mathrm{c}} \tag{4.25}$$

where $K_{P,\mathrm{c}}$ is the power gain of the frequency converter and N_{amp} is the noise figure of the amplifier. The two quantities depend on the local-oscillator voltage applied to the diode. At a low voltage, the diode current is low, and so is the converter power gain $K_{P,\mathrm{c}}$. The overall noise figure is then very high due to the second term in Eq. (4.25).

If we rise the local-oscillator voltage, $K_{P,\mathrm{c}}$ will rise, too, and the noise figure will go down. With a further increase in the local-oscillator voltage, the rate of rise in $K_{P,\mathrm{c}}$ slows down, and the second term in Eq. (4.25) stabilizes. At the same time, the rise in diode current causes an increase in the noise figure of the converter proper, that is, in the first term on the right-hand side of Eq. (4.25). The overall noise figure rises in proportion as well. It may therefore be concluded that the noise figure N has a minimum at a certain local-oscillator voltage.

In addition to shot noise and thermal noise, a frequency converter is affected by local-oscillator noise. For several causes, the phase of the voltage generated by the local oscillator is fluctuating. As a rule, these fluctuations are small, being not over a few thousandths of a degree of arc. The local-oscillator voltage is also slightly amplitude-modulated (with a modulation depth of a fraction of one percent). As a consequence, instead of a single line in its spectrum, the local-oscillator output has a spectrum containing the centre frequency f_{LO} and several side bands, as shown schematically in Fig. 4.25. In their behaviour the sidebands are not unlike the fluctuation noise of other components, and they are called local-oscillator noise. Its level is quite low, being by a factor of several hundred thousand lower than the centre-frequency level.

The width of the local-oscillator noise spectrum varies with the bandwidth of the local-oscillator tuned circuit, and it increases with increasing frequency. Considering the local-oscillator noise voltage $V_{\mathrm{n,LO}}$, the equivalent circuit shown in Fig. 4.17 may be re-drawn as shown in Fig. 4.26. As is seen, $V_{\mathrm{n,LO}}$ is combined with the signal voltage, and the sum is applied to the mixer.

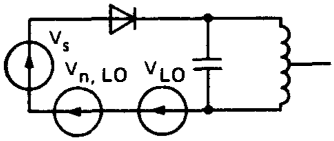


Fig. 4.26

Suppose that the noise spectrum shown in Fig. 4.25*a* contains a band with a centre frequency f_s. As a result of conversion, this frequency will be translated to f_i (it is assumed that first-order conversion $k = 1$, takes place). Falling within the bandwidth will be a portion of the noise spectrum equal to the bandwidth, B, of the i.f. amplifier. The receiver bandwidth will also contain, as a result of conversion, the local-oscillator noise band with a centre frequency equal to the image frequency, f_{im} (the two bands are shown in Fig. 4.25).

In the SHF and EHF bands, local-oscillator noise may cause an increase in the noise figure of the frequency converter by a factor of two or even more. In the UHF band and at lower frequencies, local-oscillator noise has a narrower spectrum so that f_s and f_{im} fall outside the spectrum, as shown in Fig. 4.25*b*, and local-oscillator noise does not cause any response in the receiver. A similar situation exists when a high intermediate frequency, f_i, is used, because f_s and f_{im} are then removed farther away from f_{LO}.

4.13. The Balanced Diode Frequency Converter

In Sec. 4.8 we have examined a balanced frequency converter built around transistors. As an alternative, it may use diodes, a feature which mitigates the effect of local-oscillator noise. Referring to Fig. 4.27*a*, the voltage generated by the local oscillator, *LO*, is fed to diodes *D*1 and *D*2 in phase, whereas the signal voltage is routed via a transformer, Tr_1, to the same diodes in anti-phase. Each of the two arms acts as an unbalanced converter similar to that shown in Fig. 4.22.

Since the signal voltages applied to the two arms are 180° out of phase with each other, the same phase shift exists between the i.f. currents in the diode circuits. In the primary of transformer Tr_2 these currents are in anti-phase, and the output voltage v_{if} is proportional to their difference. When the currents i_1 and i_2 are subtracted from each other, their i.f. components are in phase, and the output voltage is proportional to their sum. At the same time and similarly to v_{LO}, the local-oscillator noise voltage $v_{n,LO}$ reaches the two diodes in phase. In consequence, the i.f. components of the currents i_1 and i_2, produced as local-oscillator noise beats with f_s and f_{im}, are likewise in phase. Since the circuit is balanced and symmetrical, they cancel each other, and there is no voltage appearing at the output.

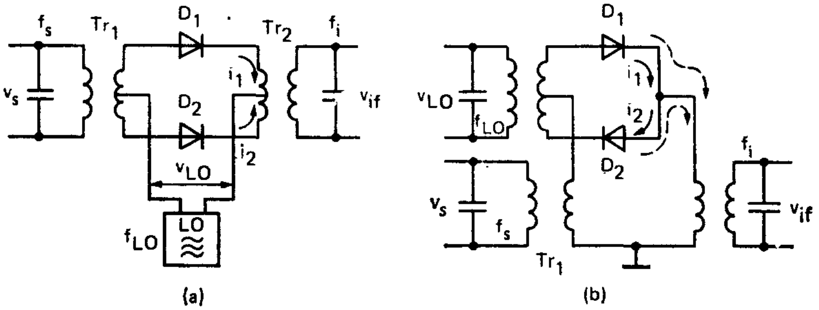


Fig. 4.27

If a receiver uses no r.f. amplifier, its frequency converter will be directly coupled to the antenna. The local-oscillator voltage, reaching the input circuit produces radiation from the antenna which is objectionable from the view-point of electromagnetic compatibility as it interferes with operation of other receivers. With a balanced frequency converter, the currents at the local-oscillator frequency flowing in the two halves of the input and output transformers are in opposition and cancel each other. That is why the local-oscillator voltage in a balanced frequency converter cannot reach the input and output circuits.

A similar behaviour is shown by the frequency converter of Fig. 4.27*a* if we interchange the local oscillator and the signal source, that is, if we feed the local-oscillator voltage via transformer Tr_1 and the signal voltage to the centre taps on the transformer windings.

In the frequency converter of Fig. 4.27*b*, the signal and local-oscillator voltages are applied each to a pair of opposite junctions of a bridge circuit formed by the secondary half-windings of transformer Tr_1 and by diodes D_1 and D_2. The current components i_1 and i_2 produced by the local-oscillator voltage flow through the diodes without dividing into the circuit which joins the opposite bridge junctions and contains the input and output tuned circuits. Therefore, as in the previous case, the local-oscillator voltage cannot reach the input and output circuits. The currents produced in the output circuit by local-oscillator noise likewise cancel out.

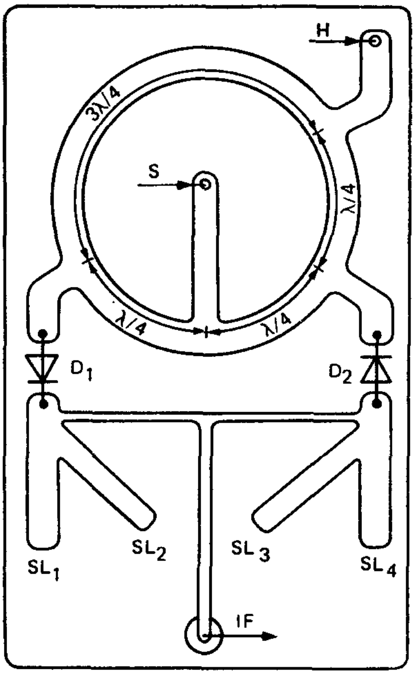


Fig. 4.28

The signal voltage is applied to diodes D_1 and D_2 in phase, but the diodes are connected back-to-back, and this produces an effect similar to that which would be produced if the voltages were in anti-phase. In the arms containing diodes D_1 and D_2, the i.f. components of currents i_1 and i_2, produced by the action of the signal, flow as shown by the dashed curves. These currents flow through the Tr_2 primary and produce an output i.f. voltage proportional to their sum.

Figure 4.28 shows the layout of a strip-line SHF balanced frequency converter. Basically, it is not unlike the frequency converter shown in Fig. 4.27*b*. Its circuits are fabricated by printed-circuit technology on the surface of a dielectric board. The surface of the board has depressions which receive two diodes, D_1 and D_2. The substrate board is covered by another dielectric board, and the opposite sides of the boards are given a coat of metal foil.

The signal and local-oscillator voltages are fed to a ring the mean circumferential length of which is one and a half times the wavelength, λ, of the signal and the local oscillator (both of them, because their frequencies differ by a very small amount equal to the intermediate frequency, f_1). The signal is applied at point *S* via a coaxial connector (not shown in the figure) which is at right angles to the plane of the drawing. Point *S* is positioned symmetrically relative to the diodes, for which reason the signal wave reaches the two diodes in phase, as it does in the circuit of Fig. 4.27.

The local-oscillator (or heterodyne) wave is injected via a coaxial connector at point *H*. The distance, as measured around the ring, from the wave-injection point to the diode D_1 tap is by a half-wave-length, $\lambda/2$, greater than the distance to the diode D_2 tap. Owing to such an arrangement, the local-oscillator wave is applied to the two diodes in anti-phase, as in the circuit of Fig. 4.27.

The heterodyned signal is picked off at point IF via a coaxial connector and goes to the i.f. amplifier. The strip lines SL_1, SL_2, SL_3 and SL_4 are each a quarter-wave, $\lambda/4$, long. An open-ended strip line of the same length acts as a series resonant circuit tuned to the signal and local-oscillator frequencies, that is, as a short-circuit. Because of this, the current at f_s and f_{LO} appearing at the outputs of D_1 and D_2 are short-circuited and do not reach the converter output.

4.14. The Capacitive Frequency Converter

The frequency converter shown in Fig. 4.18 can operate as the varactor type if a negative voltage E be applied to the diode anode. Then, on neglecting the diode conduction current, we obtain from Eq. (4.11) the following expressions for the converter parameters:

$$\left.\begin{aligned} \dot{Y}_{11} &\approx j\omega_s C_0 \\ \dot{Y}_{22} &\approx j\omega_i C_0 \\ \dot{Y}_{12} &\approx j\omega_s C_c \\ \dot{Y}_{21} &\approx j\omega_i C_c \end{aligned}\right\} \tag{4.26}$$

According to Eq. (4.13), the converter gain and its complex conjugate are

$$\left.\begin{aligned} \dot{K}_c &= \frac{j_i\omega C_c}{\dot{Y}_L + j\omega_i C_c} \\ \dot{K}_c^* &= \frac{-j\omega_i C_c}{\dot{Y}_L + j\omega_i C_c} \end{aligned}\right\} \tag{4.27}$$

If the load circuit is tuned to f_i, the reactance of that circuit along with the reactance of the diode is zero, that is

$$\left.\begin{aligned} \dot{K}_c &= j\omega_i C_c/g_L \\ \dot{K}_c^* &= -j\omega_i C_c/g_L \end{aligned}\right\} \tag{4.28}$$

For a noninverting frequency converter at resonance, we have according to Eq. (4.16)

$$\dot{Y}_{in} = j\omega_s C_0 + \omega_s \omega_i C_c^2/g_L \tag{4.29}$$

and for an inverting frequency converter we have according to Eq. (4.17)

$$\dot{Y}_{\mathrm{in}} = j\omega_s C_0 - \omega_s\omega_1 C_c^2/g_L \qquad (4.29a)$$

It is seen that an inverting varactor converter has a negative input admittance. This means that, instead of expending power from the signal source, it supplies some power by converting that which comes from the localo oscillator. The nature of this process is known from analysis of parametric phenomena in the theory of nonlinear electric circuits. The local oscillator which serves in this case not only as the modulator of diode parameters but also as a power source is called a *pump oscillator.*

The negative admittance thus obtained may be utilized for regenerative amplification of radio signals (see Secs. 1.2 and 3.9). An important feature of such amplifiers is the fact that the direct current in the varactor circuit is very low, so the shot noise level is likewise very low.

Since, as follows from Eqs. (4.29), the input conductance of a noninverting converter is positive, it cannot be used for regenerative amplification, but, as will be shown later, amplification is still possible.

Consider the behaviour of the frequency converter shown in Fig. 4.22, assuming noninverting capacitive conversion. As before, let us find V_s by Eq. (4.18) where

$$I_s = E_s g_1$$

Neglecting g_{ckt1} and assuming that the circuit is tuned to resonance, we get

$$V_s \approx E_s g_1 m_1/(g_1 m_1^2 + g_{\mathrm{in}})$$

In the SHF band, the frequency converter is ordinarily matched to the input circuit. When the signal comes from the source over a line, this arrangement prevents the occurrence of reflected waves in the line; otherwise the reflections might corrupt the received message. Also, it provides for a maximum signal voltage at the converter input, which is a desirable feature. For proper match, the input resistance of the frequency converter should be equal to the resistance of the signal source, that is,

$$g_1 = g_{\mathrm{in}}/m_1^2$$

Hence,

$$m_1 = (g_{\mathrm{in}}/g_1)^{1/2}$$

Then

$$V_s \approx E_s g_1 m_1/2g_{\mathrm{in}}$$

The output voltage is given by

$$V_{\mathrm{out}} = V_s K_m m_2$$

or, if we consider Eq. (4.28) and the expression for m_1, we have

$$V_{\text{out}} = E_{\text{s}} \times 0.5\,(g_1/g_{\text{in}})^{1/2}\, m_2\, \omega_1 C_{\text{c}}/g_{\text{L}}$$

On substituting for g_{in} from Eq. (4.29*a*) and $g_{\text{L}} = g_2 m_2^2$, the gain $K = V_{\text{out}}/E_{\text{s}}$ is found to be

$$K = 0.5\,(g_1/g_2)^{1/2} \varkappa_2$$

where

$$\varkappa_2 = (f_1/f_{\text{s}})^{1/2} \qquad (4.30)$$

In contrast to Eq. (4.24), in the case of a resistive frequency converter, one uses $\varkappa_2$ in Eq. (4.30) instead of $\varkappa_1$. As has been shown in Sec. 4.10, the factor $\varkappa_1$, which is less than unity, shows how much the conversion gain is smaller than the gain of an ideal (lossless) matching transformer. The factor $\varkappa_2$ in Eq. (4.30) may be greater than unity as well, if $f_1 > f_{\text{s}}$, which means that the converter amplifies the signal.

4.15. Regenerative Amplification

An amplifier in the form of a negative-resistance capacitive (varactor) frequency converter has an advantage over regenerative amplifiers in which negative resistance is obtained with the aid of resistive diodes (such as the tunnel diode) or feedback amplifying electron devices. Importantly, a capacitive (varactor) frequency converter has a low noise level, but it is also attractive in that its parameters can be controlled and stabilized by varying the local-oscillator voltage.

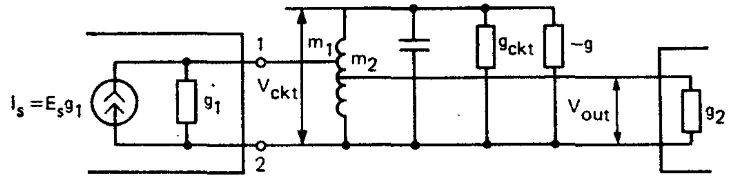


Fig. 4.29

A simplified equivalent circuit of a regenerative amplifier is shown in Fig. 4.29. Here g_1 is the signal source conductance, g_2 is the load conductance, g_{ckt} is the tuned-circuit loss conductance (the tuned circuit may be any equivalent resonator), and $-g$ is the negative conductance of the amplifying device.

Let us find the amplifier gain under the following conditions: the amplifier is matched to the signal source and its bandwidth is sufficient to accommodate the signal being amplified.

As measured between $1/\sqrt{2}$ (that is, 3-dB) points, the bandwidth is

$$B = f_0 d_{\text{eq}} = f_0 \rho g_{\text{eq}}$$

where d_{eq} is the damping factor (or attenuation) and g_{eq} is the equivalent conductance of the circuit at resonance. In the case in question,

$$g_{\text{eq}} = g_{\text{ckt}} + m_1^2 g_1 + m_2^2 g_2 - g$$

This conductance should have a certain definite value

$$g_{\text{eq}} = B_{3\text{-dB}}/f_0\rho \tag{4.31}$$

The condition for proper match between the amplifier and the signal source is

$$g_1 = (g_{\text{ckt}} + m_2^2 g_2 - g)/m_1^2$$

or, in a different way,

$$g_1 m_1^2 = g_{\text{ckt}} + g_2 m_2^2 - g \tag{4.32}$$

With a proper match,

$$g_{\text{eq}} = 2g_1 m_1^2$$

Hence,

$$m_1 = (g_{\text{eq}}/2g_1)^{1/2}$$

Noting that

$$g_1 m_1^2 = 0.5\, g_{\text{eq}}$$

we obtain from Eq. (4.27)

$$m_2 = [(0.5\, g_{\text{eq}} + g - g_{\text{ckt}})/g_2]^{1/2}$$

As follows from Fig. 4.29, we have by Ohm's law

$$V_{\text{out}} = E_s g_1 m_1\, [1/(g_1 m_1^2 + g_2 m_2^2 + g_{\text{ckt}} - g)]\, m_2$$

Subject to Eq. (4.32) and noting the expressions for m_1 and m_2, we obtain the following formula for the gain at resonance:

$$K_0 = V_{\text{out}}/E_s = 0.5\,(g_1/g_2)^{1/2}\varkappa_3$$

where

$$\varkappa_3 = [1 + 2\,(g - g_{\text{ckt}})/g_{\text{eq}}]^{1/2} \tag{4.33}$$

Since ordinarily $g_{\text{ckt}} \ll g$, it follows that

$$\varkappa_3 \approx [1 + 2\,(g/g_{\text{eq}})]^{1/2} \tag{4.34}$$

Equation (4.34) looks not unlike Eqs. (4.24) and (4.30). It likewise contains the gain of an ideal matching transformer and a factor which shows how many times the gain of the device in question differs from this gain. In the case on hand, $\varkappa_3 > 0$, which means that amplification does take place. In view of Eq. (4.31),

$$\varkappa_3 \approx [1 + 2g\,(f_0\rho/B_{3\text{-dB}})]^{1/2}$$

It is seen that one of the factors that limit amplification is the bandwidth: the wider it should be, the lower the value of $\varkappa_3$.

Another limiting factor is stability. If the negative conductance increases relative to its original value g and becomes equal to $g + \Delta g$, the equivalent conductance g_{eq} will decrease by an equal amount to become $g'_{eq} = g_{eq} - \Delta g$. With $g'_{eq} < 0$, the tuned circuit will go oscillating, and the amplifier will fail to operate as it should. The condition for no oscillations to occur is $g'_{eq} > 0$. It therefore follows that $g_{eq} > \Delta g$. In view of this condition, we obtain from Eq. (4.34)

$$\varkappa_3 \approx [1 + 2\,(g/\Delta g)]^{1/2}$$

Let the fractional change in negative conductance be denoted as $\Delta = \Delta g/g$. Then,

$$\varkappa_3 < (1 + 2/\Delta)^{1/2}$$

Or, in other words, the higher the stability (or, which is the same, the lower the value of Δ), the higher the gain that can be obtained.

4.16. Parametric Amplifier Types

An amplifier arranged as a capacitive (varactor) frequency converter is used nearly always when receiver sensitivity is to meet especially stringent requirements. In terms of noise it is somewhat inferior to a liquid-nitrogen-cooled, paramagnetic-ruby maser amplifier, for which the noise temperature is several kelvins. However, a varactor amplifier is far simpler in construction and more economical because it needs no source of a strong magnetic field which is necessary for a maser. On the other hand, masers are the devices of choice in applications where one is in a position to utilize their specific capabilities, such as radio telescopes and space communications.

A varactor amplifier ensures the lowest noise temperature in receivers which use no special cooling facilities. Its performance can be improved still more by the use of liquid nitrogen cooling and semiconductor coolers which depend for their operation on the Peltier effect. Cooling brings down the thermal noise level in receiver components. With liquid helium cooling, receiver sensitivity comes very closely to what is achieved in receivers with a maser amplifier.

Diode parametric amplifiers may be classed into the distributed-amplification type and the cascade type. An amplifier in the former type is in effect a slow-wave structure traversed by travelling waves set up by the incoming signal and the pump oscillator. Disposed along the path travelled by the wave are varactors which build up the signal energy owing to their amplifying action as the wave advances along the circuit. A *travelling-wave parametric amplifier* has a very broad bandwidth but is rather complex in construction.

The heart of a cascade parametric amplifier is a capacitive frequency converter. Parametric amplifiers in this class have a narrower bandwidth because, as has been shown in Sec. 4.15, an increase in bandwidth in the case of regenerative amplification entails a decrease in gain. In most cases, it turns out feasible to combine a sufficient gain with the required bandwidth.

In order that the noise originating in the succeeding receiver circuits could not tell on overall performance, a need may arise for consecutive signal amplification in two or three parametric amplifying stages. According to Eq. (1.26) the requirement for a low noise level is especially stringent as regards the first stage or stages. For this reason, it is customary to cool those stages.

Proceeding from the theory set forth in Secs. 4.14 and 4.15, cascade parametric amplifiers may be classed into two broad classes, namely nonregenerative and regenerative. In a nonregenerative cascade parametric amplifier the amplified signal is extracted in the i.f. circuit, the i.f. being higher than the signal frequency. As follows from Eq. (4.30), the gain depends on how many times the frequency is increased. The i.f. is the sum of the signal and local-oscillator frequencies. Regenerative cascade parametric amplifiers are essentially inverting frequency converters with a negative input conductance.

For the noise originating in the succeeding circuit not to affect the overall noise temperature of a receiver, the power gain of the amplifier should be sufficiently high, being 10 or even more. In a nonregenerative amplifier this is achieved by increasing the frequency. If the signal frequency is in the UHF band, the i.f. is chosen to lie in the SHF band. If, on the other hand, the signal frequency lies in the SHF band, the i.f. is chosen to lie in the EHF band. Since any further increase in frequency is difficult to accomplish, an amplifier usually has one stage followed by a resistive, down-converter with a gain of less than unity. The overall gain is the product of the amplifier and down-converter gains.

The amplifier may be configured as shown in Fig. 4.6*c*. In the first capacitive mixer, Mxr_1, the signal frequency is raised by a factor of 10 to 20, which fact provides for amplification. In the second resistive mixer, Mxr_2, the signal is somewhat attenuated, but still the output signal is stronger than the input signal. Since the two mixers use a common local oscillator, the construction of the device is simplified whereas the output signal frequency and phase are independent of local-oscillator stability.

A regenerative amplifier built along the lines of an inverting frequency converter may be implemented in any one of three ways as follows.

1. The amplified signal is extracted in the i.f. circuit. The intermediate frequency, $f_i = f_p - f_s$, is substantially higher than the

signal frequency, f_s. At the amplifier output the signal gains in power with increasing frequency which means that parametric amplification takes place. On the other hand, the frequency converter has a negative input conductance which means that regenerative amplification occurs as well. This type of amplifier has not found any appreciable use because it has to be operated under a complex set of conditions and its stability is anything but good.

2. The frequency converter stage is arranged as in the previous case, that is, as shown in Fig. 4.22, but use is made only of its negative input conductance. The load is coupled to the input tuned circuit which is also connected to the input of the converter (the negative conductance g in Fig. 4.29). The wave produced in the i.f. tuned circuit is not directly used, for which reason it is called the *idler frequency*, the i.f. circuit is the idler circuit, and what we have is a double-tuned-circuit parametric amplifier.

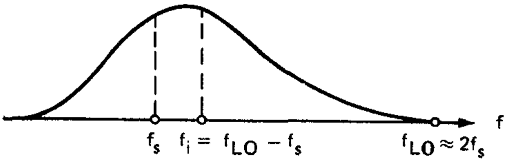


Fig. 4.30

3. The pump frequency, f_p, is close to twice the signal frequency, $2f_s$; the difference frequency $f_i = f_p - f_s$ is close to f_s and falls within the passband of the input tuned circuit, as shown in Fig. 4.30. Now there is no need for a separate output tuned circuit, and the two tuned circuits may be combined (as in, say, Fig. 4.22). In such a case, the amplifier is arranged as shown in Fig. 4.29 where the negative-conductance ($-g$) element is a varactor driven by a pump. What we have is a *single-tuned-circuit regenerative parametric amplifier.*

Frequencies f_s and f_i (see Fig. 4.30) ought not to lie too close to each other, or else it would be difficult to separate them after the succeeding down-conversion in the i.f. amplifier.

In contrast to the case shown in Fig. 4.30, f_s and f_i in real amplifiers are bands of frequencies rather than single frequencies. If the two bands overlap, then, after one of these spectra has been selected, some portions of the other, falling within the passband of the i.f.. amplifier, will corrupt the desired signal, thus acting as interference.

A special thing about this case is that the current at the converted frequency flows through an only tuned circuit to which are connected the transferred source and load conductances. In contrast to the circuit in Fig. 4.22, the equivalent load for the converter will be

$$g_{L,eq} = g_1m_1^2 + g_2m_2^2$$

This quantity should enter the denominator in Eq. (4.29*a*) for the negative conductance of the varactor.

Since, on top of that, $\omega_1 \approx \omega_s$, the negative conductance will be

$$g \approx -\omega_s^2C_c^2/g_{L,eq} \qquad (4.35)$$

The condition for proper match between the amplifier and the signal source, Eq. (4.32), has the same form as before, that is, neglecting the tuned-circuit loss conductance, g_{ckt}

$$g_1m_1^2 = g_2m_2^2 - g = 0.5g_{eq} \qquad (4.36)$$

The value of m_1 may be chosen, using the same equation as was used in Sec. 4.15, that is,

$$m_1 = (g_{eq}/2g_1)^{1/2}$$

The situation is somewhat different with the choice of m_2. Noting the expressions for g and $g_{L,eq}$, we obtain from Eq. (4.36)

$$g_{eq}m_2^2 - g_1m_1^2 = (\omega_s^2C_c^2)/(g_2m_2^2 + g_1m_1^2)$$

Hence,

$$(g_2m_2^2)^2 - (g_1m_1^2)^2 = \omega_s^2C_c^2$$

or, in a different way,

$$(g_2m_2^2)^2 = \omega_sC_c^2 + (0.5g_{eq})^2$$

wherefrom

$$m_2 = \{[(\omega_sC_c)^2 + (0.5g_{eq})^2]/g_2^2\}^{1/4}$$

Also, as follows from Eq. (4.35)

$$g = \omega_s^2C_c^2/\{0.5\,g_{eq} + [\,\omega_s^2C_c^2 + (0.5g_{eq})^2]\,\}^{1/2}$$

If we denote

$$\alpha = 2\,\omega_sC_c/g_{eq} \qquad (4.37)$$

we may then re-write the last expression as

$$g = \omega_sC_c\alpha/[1 + (1 + \alpha^2)]^{1/2}$$

Substituting the above expression for g in Eq. (4.34) gives

$$\varkappa_3 \approx (1 + \alpha^2)^{1/4}$$

It is seen that the gain increases with increasing C_c and decreases with increasing g_{eq}, that is, with increasing bandwidth.

4.17. Sources of Noise in the Single-Tuned-Circuit Parametric Amplifier

Noise levels in a single-tuned-circuit parametric amplifier can be calculated on the basis of the simplified equivalent circuit which is shown in Fig. 4.31 and which corresponds to that in Fig. 4.29. Here, g_1 and g_2 are the source and load conductances transferred into the resonant circuit, whereas I_1, I_2, I_3 and I_4 are noise currents.

Current I_1 is produced by the signal source. In finding the noise figure, it is assumed, as follows from the procedure set forth in Secs. 1.7 and 3.8, to be proportional to the power arriving at the input from the matched source. The amplifier is assumed to be matched to the signal source, which means that the condition defined by Eq. (4.32) is satisfied. Therefore, the noise due to the $m_1^2g_1$ branch satisfies the conditions for calculating the noise figure. According to Eq. (1.5), the mean square of this current is

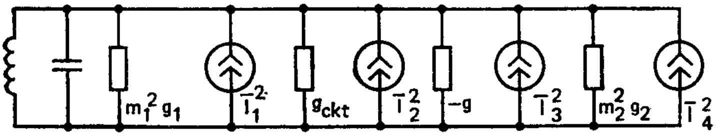


Fig. 4.31

$$\overline{I_1^2} = 4kTBg_1m_1^2$$

The loss occurring in the tuned circuit proper is associated with the noise current

$$\overline{I_2^2} = 4kTBg_{\text{ckt}}$$

As the theory of electron devices tells us, the mean square of diode noise current, mainly determined by shot noise, is

$$\overline{I_3^2} \approx eI_0B$$

where e is the electron charge and I_0 is the direct component of diode current. For uniformity, this current may alternatively be written as

$$\overline{I_3^2} = 4kTBg_{\text{n,D}}$$

where $g_{\text{n,D}}$ is an arbitrary quantity having the dimensions of conductance. It is equivalent to diode noise and depends on I_0 and some other factors which come into play under concrete service conditions. Taking the equality

$$2eI_0B = 4kTBg_{\text{n,D}}$$

we find

$$g_{n,D} = eI_0/2kT$$

but it may be verified and refined by testing the diode in an experiment.

The noise current I_4 arises in the input circuit of the next stage and may exceed the input-conductance thermal current of this amplifier, as found by Eq. (1.5) at room temperature. Taking this fact into account, we may write the mean square of this current as

$$\overline{I_4^2} = 4k\gamma TBm_2^2g_2$$

where the value of γ depends on the type and properties of the next stage.

The noise output power is proportional to the mean square of the noise voltage across the tuned circuit and, as a consequence, to the mean square of noise current. Therefore the amplifier noise figure may be written as

$$N_{\text{amp}} = (\overline{I_1^2} + \overline{I_2^2} + \overline{I_3^2} + \overline{I_4^2})/\overline{I_1^2}$$

On substituting for the respective currents, we get

$$N_{\text{amp}} = 1 + [(g_{\text{ckt}} + g_{n,D} + \gamma m_2^2g_2)/m_1^2g_1]$$

Subject to the condition for match, Eq. (4.32), we may write

$$N_{\text{amp}} = 1 + [g_{\text{ckt}} + g_{n,D} + \gamma\ (0.5g_{\text{eq}} + g - g_{\text{ckt}})]/0.5g_{\text{eq}}$$

or, in a different way,

$$N_{\text{amp}} = 1 + 2\ (g_{\text{ckt}} + g_{n,D})/g_{\text{eq}} + \gamma\ [1 + 2\ (g - g_{\text{ckt}})/g_{\text{eq}}] \quad (4.38)$$

The term in the square brackets in Eq. (4.38) is the power gain, K_P. To demonstrate, as is seen from Fig. 4.29, the power applied from the signal source to the amplifier is

$$P_{\text{in}} = V_{\text{ckt}}^2\ (g_{\text{ckt}} - g + m_2^2g_2)$$

and the power transferred to the load is

$$P_{\text{out}} = V_{\text{ckt}}^2m_2^2g_2$$

From the condition for match, Eq. (4.36), it follows, however, that

$$m_2^2 = 0.5g_{\text{eq}} + g - g_{\text{ckt}}$$

and

$$g_{\text{ckt}} - g + m_2^2g_2^2 = 0.5g_{\text{eq}}$$

Consequently,

$$P_{\text{out}}/P_{\text{in}} = K_P = (0.5g_{\text{eq}} + g - g_{\text{ckt}})/0.5g_{\text{eq}}$$

Therefore, Eq. (4.38) may be re-cast as

$$N_{\text{amp}} = 1 + \gamma K_P + 2\ (g_{\text{ckt}} + g_{n,D})/g_{\text{eq}}$$

The noise figure for the amplifier together with the succeeding stages for which the noise figure is N_s can be found by Eq. (1.26) which, given a proper match between the stages, takes the form

$$N = N_{amp} + (N_s - 1)/K_P$$

It is seen from Eq. (4.38) that the noise figure may rise due to the noise current I_4 associated with the factor γ. The strong effect produced by this current stems from the fact that in an amplifier the input and the output are coupled. Noise from the output circuit goes to the amplifier and is amplified along with the wanted signal. To prevent output-circuit noise from being amplified, it is necessary to arrange so that it has no effect on the amplifier. Most commonly this task is achieved by coupling the input and output circuits to the amplifier via circulators.

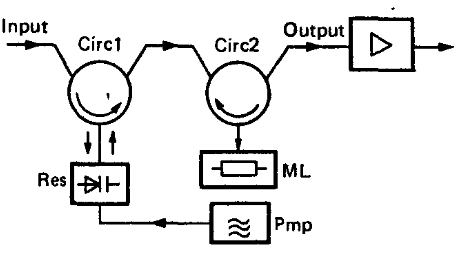


Fig. 4.32

In simplified form the arrangement of an amplifier using circulators is shown in Fig. 4.32. The incoming signal is routed to a resonator, *Res*, enclosing a varactor, via a circulator, *Circ*$_1$. The resonator is also fed the wave from a pump oscillator (or, simply, a pump), *Pmp*. The same waveguide channels the signal back to *Circ*$_1$ from which it goes to a second circulator, *Circ*$_2$, leaves it to appear at the output of the amplifier, and proceeds to the input of the next amplifier. Noise and the reflected waves which may go back from the output to the second circulator cannot reach the amplifier; rather, they are absorbed by a matched load, *ML*. Thus, the circulators turn the amplifier into a nonreciprocal device, and amplification can therefore be effected consecutively in several stages.

Figure 4.33 shows the circuit of a two-stage amplifier in which the first stage is cooled. The incoming signal is routed by a waveguide, W_1, to the first circulator, *Circ*$_1$, and then proceeds to the first stage of a parametric amplifier, *PA*$_1$, which also receives a signal from a pump, *Pmp*, via a second waveguide, W_2. The amplified signal is transferred via a third waveguide, W_3, and the first filter, *Filt*$_1$, to the second circulator, *Circ*$_2$. The filter passes the wanted signal and blocks the wave travelling down W_3 from the pump and also spurious-response noise. From the second circulator, the signal proceeds through a similar second filter, *Filt*$_2$, and a third circulator, *Circ*$_3$, to the second amplifier stage. The pump feeds this stage via a fourth waveguide, W_4. The signal amplified by the second stage is transferred via a third filter, *Filt*$_3$, and a fourth circulator, *Circ*$_4$, goes to a down-converter which consists of a diode mixer, *Mxr*, and a local oscillator, *LO*. On leaving the mixer, the heterodyned signal goes to an i.f. amplifier. The waves reflected by $Circ_1$ and $Circ_4$ are absorbed by matched loads, *ML*.

In some designs, there may be three, four or even more amplifier stages, with some of the early stages being cooled. To achieve the ultimate in noise reduction, the first stage and the associated circuits, including the first circulator, are enclosed in a double Dewar flask filled with liquid helium, He, and liquid nitrogen, N.

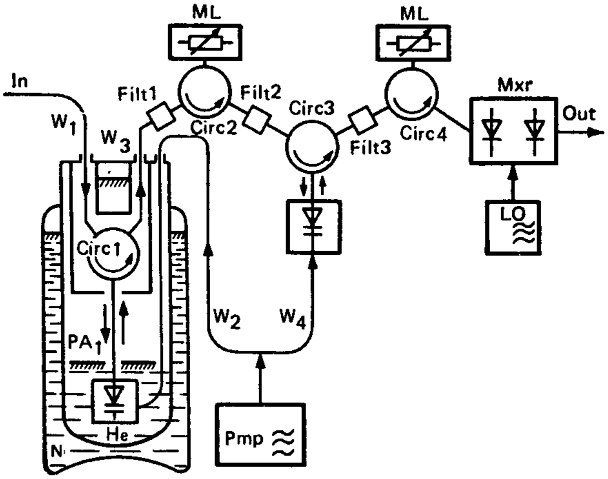


Fig. 4.33

4.18. The Gain of an Amplifier with a Circulator

Since in an amplifier with a circulator (see Fig. 4.32) the load has no back-effect on the current from the signal source, the gain is greater than it is for the regenerative amplifier examined in Sec. 4.15, Eq. (4.34). With a perfect match at the amplifier input, the input voltage is, as in the cases examined earlier, given by

$$V_{\text{in}} = 0.5E_s\,(g_1/g_2)^{1/2}$$

where g_2 is now the input (characteristic) conductance of the waveguide coupling the signal source to the first circulator. Associated with this voltage is the input current

$$I_{\text{in}} = V_{\text{in}}g_2$$

which is also fed to the resonator, *Res*. Considering the losses in the waveguide and circulator, this current is equal to $E_s \times 0.5\,(g_1g_2)^{1/2}\beta_1$, where $\beta_1 < 1$.

Assume that the input conductances of all the waveguides feeding the circulators are equal to g_2. If the waveguide running from the circulator is connected to the resonator so that the tapping-down factor is m, and if we neglect the losses inherent in the resonator, the resonator conductance at resonance will be

$$g_{eq} = g_2 m^2 - g$$

Therefore, assuming the bandwidth corresponding to g_{eq}, we have

$$m = [(g_{eq} + g)/g_2]^{1/2}$$

The voltage produced across the resonator by I_{in} is

$$V_2 = I_{in} m/g_{eq}$$

The voltage fed back to the first circulator is mV, whereas the voltage existing at the output of the entire amplifier, that is, at the output of the second circulator is $mV\beta_2$, where β_2 takes care of the further losses in the two circulators and connecting waveguides. Therefore, in contrast to Sec. 4.15, we now have

$$V_{out} = 0.5\,(g_1 g_2)^{1/2}\beta_1 m^2 \beta_2/g_{eq}$$

On substituting for m, the gain at resonance is

$$K_0 = V_{out}/E_s = 0.5\,(g_1/g_2)^{1/2}\beta_1\beta_2\,(1 + g/g_{eq})$$

or, in a different way,

$$K_0 = 0.5\,(g_1 g_2)^{1/2}\varkappa_4$$

where

$$\varkappa_4 = \beta_1\beta_2\,(1 + g/g_{eq})$$

On introducing, as we did in Sec. 4.15, the condition for an amplifier not to go oscillating, we may write

$$\varkappa_4 = \beta_1\beta_2\,(1 + \Delta)$$

Where use is made of low-loss circuits, the factors β_1 and β_2 are slightly less than unity. Comparison of $\varkappa_4$ and $\varkappa_3$ shows that owing to a circulator the gain markedly increases (in the absence of losses it nearly doubles).

Chapter Five

Radio-Signal Detectors

5.1. Detector Types and Key Characteristics of Amplitude-Modulation Detectors

A *detector* (also called a *demodulator*) is a device which operates on a modulated carrier wave to recover the wave with which the carrier was originally modulated. Accordingly to the type of modulation used, there may be amplitude, frequency and phase detectors (or demodulators).

Amplitude-modulation (AM) detection can be effected by nonlinear circuits (NE in Fig. 5.1*a*) and by synchronous detectors. Being simpler in design, nonlinear AM detectors are the predominant type.

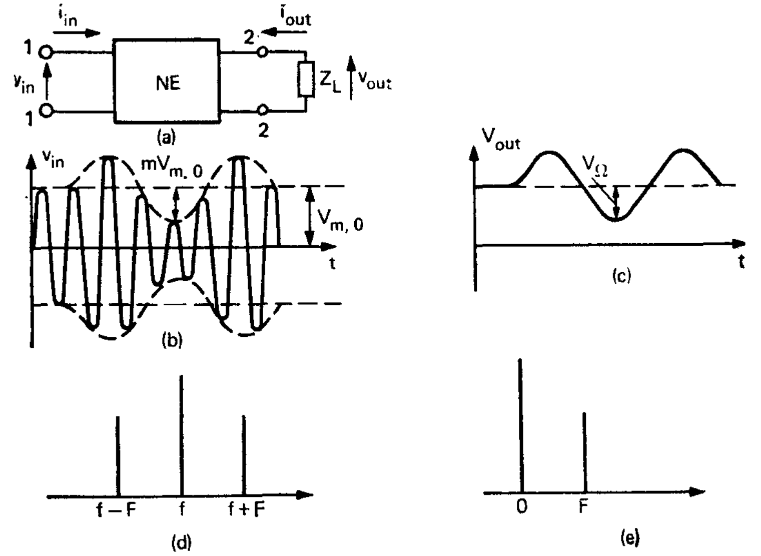


Fig. 5.1

Let, at the input, there be a single-tone amplitude-modulated voltage (Fig. 5.1*b*)

$$V_{in} = V_{m,0}\,(1 + m\cos\Omega t)\cos\omega t \tag{5.1}$$

Its spectrum is shown in Fig. 5.1*d*.

The voltage developing across the detector load contains a d.c. component and an a.c. component of the form shown in Fig. 5.1*c*. The useful result of detection is the component

$$V_{out} = V_\Omega\cos\Omega t \tag{5.2}$$

The spectrum of the detector output voltage is shown in Fig. 5.1*e*.

Synchronous detection is effected as the signal defined by Eq. (5.1) is multiplied by a reference voltage

$$v_{ref} = V_{ref}\cos\omega t$$

The resultant voltage

$$v_{out} = A\,(1 + m\cos\Omega t)\,(0.5 + 0.5\cos 2\omega t)$$

contains a component at frequency 2ω which is suppressed by a low-pass filter. The remaining component contains the useful result of detection of the form given by Eq. (5.2).

Synchronous-detection circuits are not unlike those used for frequency conversion, except that the job of the local oscillator is done by the reference voltage and the i.f. filter at the output is replaced by a low-pass filter.

Detection may cause distortion, both linear and nonlinear, in the signal. Nonlinear distortion is evaluated in terms of what we have termed the harmonic factor*

$$k_h = (V^2_{2\Omega} + V^2_{3\Omega} + \ldots)^{1/2}/V_\Omega$$

where $V_{2\Omega}$, $V_{3\Omega}$, etc. are the amplitudes of output voltage at angular frequencies 2Ω, 3Ω, etc.

Linear (that is, amplitude and phase) distortion is due to the fact that any amplitude-modulation detector contains inertial elements, mainly capacitances. Amplitude distortion (sometimes called frequency distortion) arises from the fact that the detector gain K varies with the modulation frequency, Ω, of the incoming signal. Phase (or delay) distortion occurs when the phase shift of the output voltage with respect to the envelope of the input radio signal is not linearly related to the modulation frequency.

As has been noted in Sec. 1.5, the detector gain is defined as the ratio of the amplitude of output voltage, V_Ω to the amplitude of the envelope of input modulated voltage $mV_{m,0}$:

$$K_d = V_\Omega/mV_{m,0}$$

The ex'ent to which a detector affects the associated signal source is characterized by the detector input admittance. Owing to the resonant properties of the signal source, this effect is mainly determined by the fundamental component of the input current. The input admittance is defined as the ratio of the amplitude of the input current fundamental component, $\dot{I}_\omega$, to the amplitude of the signal carrier voltage at the detector input

$$\dot{Y}_{in} = \dot{I}_\omega/\dot{V}_{m,0}$$

It has a real part, G_{in}, and an imaginary part, $j\omega C_{in}$, such that

$$Y_{in} = G_{in} + j\omega C_{in}$$

The imaginary, or capacitive, part is balanced out by appropriately tuning the resonant circuit. Then the input admittance may be treated as a pure conductance.

5.2. Types of Amplitude-Modulation Detectors

The nonlinear element used in an AM detector may be a diode, a transistor, or an IC. The most commonly used type is the *diode detector*. It is simple in construction and gives an almost distortion-

* See the footnote on page 41.— *Translator's note.*

free detection over a wide range of signal levels. There may be a *series diode detector* and a *shunt diode detector*. Their schematics are shown in Figs. 5.2*a* and *b*, respectively. Both operate on the same principle. An advantage of the shunt diode detector is that there is no conductive (or resistive) coupling between the signal source and the diode.

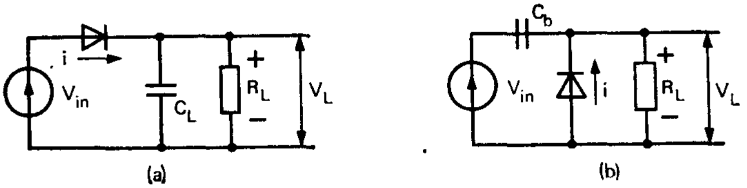


Fig. 5.2

Consider a series diode detector, assuming to a first approximation that the ideal is a ideal device, that is, one with a linear, unidirectional (or unilateral) characteristic. When an input voltage is applied, current pulses flow through the diode (Fig. 5.3). These pulses contain a direct component, I_L, and components at angular frequencies ω, 2ω, etc. The direct component produces the load voltage

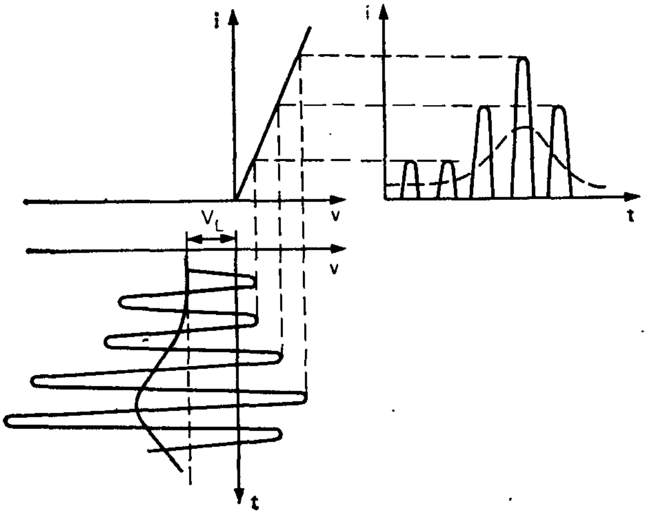


Fig. 5.3

$$V_L = -I_L R_L$$

whereas the h.f. components have their path completed through a capacitor, C_L, which presents a very low reactance at such frequencies. In amplitude modulation, there are changes in the amplitude of pulse currents and, in consequence, in their mean value and in the voltage across R_L. For the current at the modulation frequency to flow through R_L and for the currents at frequencies ω, 2ω, etc. to flow via C_L, it is essential to satisfy the following inequalities:

$$(\omega C_L)^{-1} \ll R_L \ll (\Omega_h C_L)^{-1} \tag{5.3}$$

where Ω_h is the upper modulation frequency.

In a shunt diode detector, the voltage developing across R_L will also contain an alternating voltage v_{in} in addition to the detected (rectified) voltage. If this alternating voltage is not to find its way into the succeeding circuits, one should either insert a low-pass filter or pick the detected voltage off the d.c. blocking capacitor C_b.

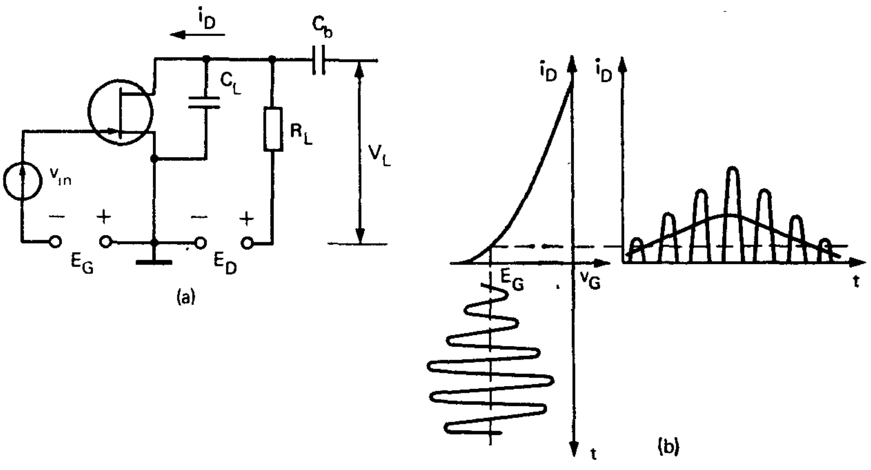


Fig. 5.4

AM detectors which use an amplifying device for their nonlinear element effect both detection and amplification. Figure 5.4*a* shows a FET detector in which the load is placed in the drain lead (for which reason it is sometimes called a drain FET detector). Detection is effected owing to the nonlinear transfer characteristic, $i_D = \varphi(v_G)$, of the FET (see Fig. 5.4*b*). The initial bias voltage at which the FET is near cutoff is supplied by a power supply in the gate lead, E_G. When a signal voltage, V_{in}, is applied to the input, current pulses begin to flow in the drain circuit. The detected current slowly varying at the modulation frequency gives rise to a voltage across the load resistor, R_L. Current components at angular frequencies ω, 2ω, etc. have their path completed through C_L. This type of detector has a high input impedance.

A transistor detector may have its load placed in the collector, base, or emitter lead. Accordingly, there may be a collector-load transistor detector, a base-load transistor detector, and an emitter-load transistor detector. Figure 5.5 shows the circuit of a collector-load transistor detector in which detection is effected owing to the nonlinear transfer characteristic, $i_C = \varphi(V_{BE})$ of the transistor. For proper operation of the detector, the time constant R_1C_1 is chosen such that

$$(\omega C_1)^{-1} \ll R_1 \ll (\Omega_h C_1)^{-1}$$

The detection taking place in the base and collector circuits of the transistor is opposite in its effect on the collector current. This causes a fall in the gain, but minimizes nonlinear distortion and increases the maximum amplitude of the input signal at which the collector circuit does not still produce a limiting effect. This is what may be called a collector-base detector.

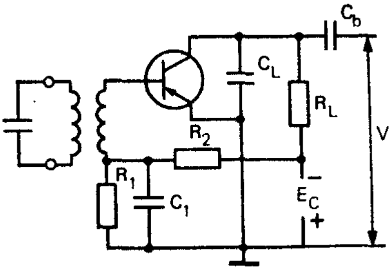


Fig. 5.5

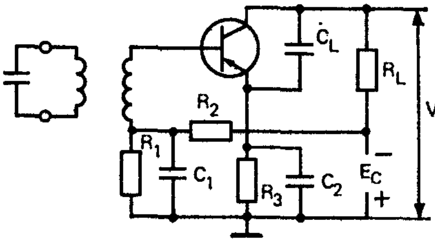


Fig. 5.6

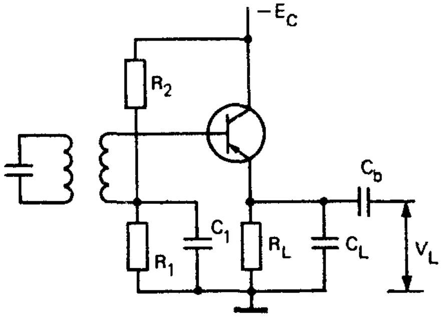


Fig. 5.7

The detection characteristic may further be linearized at the expence of a reduction in gain by applying negative envelope feedback. To this end, the emitter circuit is extended to include a resistor-capacitor network (R_3C_2 in Fig. 5.6) with a time constant chosen such that the current components at the carrier and its harmonics have their path completed through C_2, whereas the current at the modulation frequency produces a voltage drop across R_3 and, as a consequence, negative feedback.

In the emitter detector of Fig. 5.7, the time constant R_LC_L is chosen so as to satisfy the inequalities defined in (5.3). Detection is effected owing to the nonlinear transfer characteristic $I_E = \varphi(V_{BE})$ of the transistor. This type of detector uses a nearly 100% negative envelope feedback. Owing to such an arrangement, there is no overloading by large signals and the input impedance is high, but the detector gain is less than unity.

5.3. Weak-Signal Detection Theory

A detector may be depicted as a nonlinear two-port loaded into an impedance, Z_L, as shown in Fig. 5.1*a*. The input voltage is

$$v_{in} = V_m(t) \cos \omega t$$

Among the components of input current i_{in}, we are interested in the component I_ω at angular frequency ω, because it determines the input impedance. Among the components of output current i_{out}, we are interested in its slowly varying component I_L which produces the useful result of detection. Assume that the nonlinear element is free from inertia. Then the input and load currents of the detector will be functions of only the applied voltages

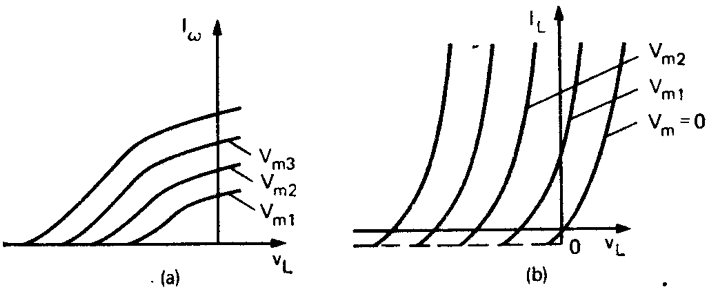


Fig. 5.8

$$I_\omega = \varphi_1(V_m, V_L) \tag{5.4}$$
$$I_L = \varphi_2(V_m, V_L)$$

where φ_1 and φ_2 are functions whose form depends on the properties of the nonlinear two-port. The functions of the form φ_1 are plotted as oscillatory characteristics (Fig. 5.8*a*), and the functions of the form φ_2 are plotted as rectification characteristics (Fig. 5.8*b*).

Since a detector is a nonlinear circuit, its properties are markedly dependent on the voltage of the signal being detected. Whereas in an amplifier, nonlinearity is objectionable as it distorts strong signals, it is desirable in a detector at any signal level because it provides the very basis of signal detection. In fact, as the theory of non-linear circuits shows, the detected voltage is a more linear function of the input signal amplitude in the case of strong signals than it is in the case of weak signals. Thus, when a detector handles strong signals, their distortion is reduced, which is why it is advantageous to apply relatively strong signals to a detector. For diode detectors, weak signals are those with V_m less than 0.25 volt. For transistor detectors, V_m less than 25 mV.

Let us write the volt-ampere characteristic of a non-linear element as

$$i = \varphi\,(E + v) \tag{5.5}$$

where E is the initial bias voltage which may be equal to zero in a particular case, such as shown in Fig. 5.2. Consider a diode detector for which, according to Fig. 5.2*a*,

$$v = V_m \cos \omega t - V_{\mathrm{L}}$$

where

$$V_{\mathrm{L}} = I_{\mathrm{L}} R_{\mathrm{L}}$$

With weak input signals, the detected signal will likewise be weak, therefore Eq. (5.5) may be expanded into a Taylor series

$$i = \varphi\,(E) + \varphi'\,(E)\,v + 0.5\varphi''\,(E)\,u^2 + \ldots \tag{5.6}$$

where $\varphi\,(E)$ is the no-signal current close to zero at $E = 0$. In the subsequent discussion, we will limit ourselves to this case. Let us denote

$$S = \varphi'\,(E) \text{ and } S' = \varphi''\,(E) \tag{5.7}$$

On substituting for v and inserting the quantities defined by Eq. (5.7) in Eq. (5.6), we re-arrange it to obtain

$$i = (S - S'V_{\mathrm{L}})\,V_m \cos \omega t - SV_{\mathrm{L}} + 0.5S'V_{\mathrm{L}}^2 + 0.25S'V_m^2 + 0.25S'V_m^2 \cos 2\,\omega t + \ldots \tag{5.8}$$

Hence, the amplitude of current at frequency ω and the direct component are

$$I_\omega = (S - S'V_{\mathrm{L}})\,V_m \tag{5.9}$$

$$I_{\mathrm{L}} = -SV_{\mathrm{L}} + 0.25S'V_{\mathrm{m}}^2\,[1 + 2\,(V_{\mathrm{L}}/V_m)^2] \tag{5.10}$$

In the detection of weak signals,

$$(V_{\mathrm{L}}/V_m)^2 \ll 1$$

Therefore, on substituting $V_{\mathrm{L}} = I_{\mathrm{L}} R_{\mathrm{L}}$ in Eq. (5.10), we get

$$I_{\mathrm{L}} = [0.25S/(1 + SR_{\mathrm{L}})]\,V_m^2 = AV_m^2 \tag{5.11}$$

As is seen, the detector has a square-law characteristic. If the signal amplitude varies as

$$V_m = V_{m.0}\,(1 + m \cos \Omega t)$$

then, according to Eq. (5.11)

$$I_{\mathrm{L}} = AV_{m,0}^2\,(1 + 2m \cos \Omega t + 0.5m^2 + 0.5m^2 \cos 2\Omega t)$$

In addition to the component at the modulation frequency $I_\Omega = 2AmV^2_{m,0}$, the detector current contains the second harmonic

$$I_{2\Omega} = 0.5Am^2V^2_{m,0}$$

responsible for nonlinear distortion such that

$$k_{\mathrm{h}} = I_{2\Omega}/I_\Omega = 0.25m$$

The detector gain

$$K_{\mathrm{d}} = V_\Omega/mV_{m,0} = I_\Omega R_{\mathrm{L}}/mV_{m,0} = 2AR_{\mathrm{L}}V_{m,0}$$

is a function of the signal carrier amplitude. Since a weak signal is being detected, the detector gain is low. In view of the above shortcomings, weak-signal detection is not used in most receivers.

5.4. Diode Detection of Strong Signals

Examine the principle by which the series diode detector shown in Fig. 5.2a operates. Strong signals are detected with current cutoff, as shown in Fig. 5.3. In the general case, the diode has a forward and a reverse current branch in its volt-ampere characteristics. To begin with, consider the operation of the diode without reverse current. To a first approximation, the forward current characteristic in the case of strong signals may be replaced by a straight line. Then the idealized diode characteristic will have the form

$$i = \begin{cases} Sv & \text{for } v > 0 \\ 0 & \text{for } v \leqslant 0 \end{cases} \tag{5.12}$$

where S is the slope of the diode characteristic more commonly referred to as the mutual conductance or transconductance.*

When an unmodulated signal is applied, such that

$$v_{\mathrm{in}} = V_m \cos \omega t$$

the voltage across the diode will be

$$v = V_m \cos \omega t - V_{\mathrm{L}} \tag{5.13}$$

As is seen from Fig. 5.9, the current appears as pulses with a cutoff angle θ** such that

$$v = V_m \cos \theta - V_{\mathrm{L}} = 0$$

whence

$$\cos \theta = V_{\mathrm{L}}/V_m \tag{5.14}$$

* Quite often, the symbol for this quantity, especially outside the USSR, is g_m.— *Translator's note.*

** It is common outside the USSR to use the concept of the angle of current flow (in the U. K.), and the operating or conduction angle (in the USA), which is twice the cutoff angle θ.— *Translator's note.*

In view of Eq. (5.14), we may re-write Eq. (5.13) as

$$v = V_m\,(\cos\,\omega t - \cos\,\theta) \tag{5.15}$$

Hence, according to Eq. (5.12)

$$i = \begin{cases} SV_m\,(\cos\,\omega t - \cos\,\theta) & \text{for } \omega t < \theta \\ 0 & \text{for } \omega t > \theta \end{cases} \tag{5.16}$$

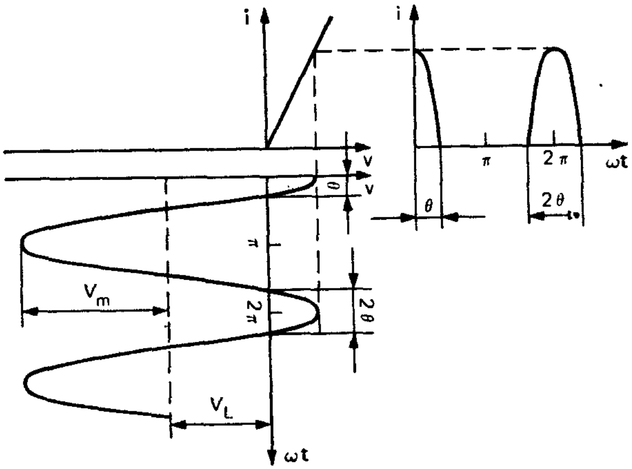


Fig. 5.9

The current contains a direct component, I_L, a component at frequency ω, and its harmonics. The result produced by detection is given by

$$I_L = (1/\pi)\int_0^\theta i(\omega t)\,dt$$

$$= (1/\pi)\int_0^\theta SV_m\,(\cos\omega t - \cos\theta)\,d\omega t$$

$$= (SV_m/\pi)\,(\sin\theta - \theta\cos\theta) \tag{5.17}$$

In order to find the cutoff angle, multiply both sides of Eq. (5.17) by R_L

$$V_L = (SR_L/\pi)\,V_m\,(\sin\,\theta - \theta\,\cos\,\theta) \tag{5.18}$$

or, on substituting (5.14) in (5.18),

$$\tan\,\theta - \theta = \pi/SR_L \tag{5.19}$$

It is seen that the cutoff angle θ is a constant quantity, which means that the load current I_L and the load voltage V_L are linear functions of the input signal amplitude. In other words, a strong-signal detector has a linear response. In this sense, it is called a linear detector.

In the general case, the transcendental equation (5.19) has no analytic solution. At low values of θ, it may be taken that

$$\tan\theta \approx \theta + \theta^3/3$$

in which case we obtain from Eq. (5.19)

$$\theta = (3\pi/SR_L)^{1/3} \tag{5.20}$$

The cutoff angle θ has a direct bearing on all the key parameters of a detector.

The value of C_L is chosen such that detection produces no frequency distortion. Then, if

$$V_m = V_{m,0}\,(1 + m\cos\Omega t)$$

then the output voltage can, according to Eq. (5.14), be found as

$$V_L = V_{m,0}\cos\theta\,(1 + m\cos\Omega t)$$

The amplitude of the alternating component of output voltage is

$$V_\Omega = mV_{m,0}\cos\theta$$

In consequence, the detector gain (see Sec. 5.1) is

$$K_d = V_\Omega/mV_{m,0} = \cos\theta \tag{5.21}$$

The amplitude of the input current fundamental can be found on expanding Eq. (5.15) into a Fourier series

$$\begin{aligned} I_\omega &= (2/\pi)\int_0^\theta i(\omega t)\cos\omega t\,d\omega t \\ &= (2/\pi)\int_0^\theta SV_m(\cos\omega t - \cos\theta)\cos\omega t\,d\omega t \\ &= (SV_m/\pi)(\theta - \sin\theta\cos\theta) \end{aligned} \tag{5.22}$$

Hence, the input conductance of the detector is

$$G_{in} = I_\omega/V_m = (S/\pi)(\theta - \sin\theta\cos\theta) = (S/\pi)(\theta - 0.5\sin 2\theta) \tag{5.23}$$

At low values of θ, we may use the expansion of the form $\sin\alpha \approx \approx \alpha - \alpha^3/6$ and, recalling Eq. (5.20), we obtain from Eq. (5.23)

$$G_{in} \approx 2/R_L \tag{5.24}$$

At low values of θ (such that $\cos\theta \approx 1$), Eq. (5.24) can be derived, proceeding from the fact that practically all of the signal power applied to the detector is dissipated in the load resistor, $P_{in} \approx P_L$.

Then

$$V_m^2/2R_{in} \approx V_L^2/R_L \tag{5.25}$$

According to Eq. (5.14),

$$V_L = V_m \cos\theta$$

At $\cos\theta \approx 1$, $V_L \approx V_m$, and Eq. (5.25) leads to Eq. (5.24).

In a shunt diode detector (see Fig. 5.2*b*), the input conductance is equal to the sum of the load and diode conductances, Eqs. (5.23) and (5.24):

$$G_{in,shunt} = G_{in} + 1/R_L \approx 3/R_L \tag{5.26}$$

As is seen, it is greater than the input conductance of a series diode detector.

It is not always that the reverse current may be neglected when a detector uses a germanium diode. Owing to its effect, the diode acquires an infinite reverse transconductance

$$1/R_r = S_r$$

as shown in Fig. 5.10. It brings about a change in the equivalent load resistance and in the input conductance. The d.c. equivalent load resistance of such a detector is

$$R_{L,eq} = R_L R_r/(R_L + R_r) \tag{5.27}$$

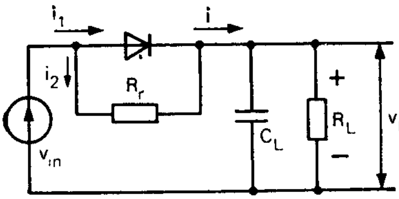


Fig. 5.10

When R_r is many times R_L, then $R_{L,\,eq}$ is almost equal to R_L.

The input conductance at $\cos\theta \approx 1$ can be found, proceeding from the fact that the input power of the detector is nearly equal to its output power:

$$P_{in} \approx P_L + P_r$$

Hence,

$$V_m^2/2R_{in} \approx V_L^2/R_{L,eq} + V_m^2/2R_r \tag{5.28}$$

With $\cos\theta \approx 1$, $V_L \approx V_m$. Therefore, as follows from Eq. (5.28) and subject to Eq. (5.27), we have

$$G_{in} \approx 2/R_{L,eq} + 1/R_r = (3R_L + 2R_r)/R_L R_r \tag{5.29}$$

The reverse transconductance of a diode increases the input conductance of the detector. When R_r is many times R_L, we have

$$G_{in} \approx 2/R_L$$

5.5. Distortion in Diode Detection of Strong Signals

As has been noted, in the case of strong signals the detector characteristic is close to linear. Distortion will be negligible, if the signal amplitude does not fall below some value which usually is

$V'_m = 0.05$ to 0.1 volt. In consequence, in order to avoid nonlinear distortion due to a decrease in the signal amplitude at high values of the amplitude modulation factor, it is essential to satisfy any one of two conditions

$$V_{m,0}(1-m) \geqslant V'_m$$

or

$$V_{m,0} \geqslant V'_m (1-m)^{-1}$$

For example, at $m = 0.9$, the carrier amplitude $V_{m,0}$ should be in excess of 0.5 to 1 volt.

However, nonlinear distortion may arise even with strong signals owing to the inertia of the detector load and the difference in load resistance to direct and alternating current.

The effect of load inertia is illustrated in Fig. 5.11. During the positive half-cycles of input voltage, the diode is conducting, and C_L is charged through its low resistance, as shown in Fig. 5.2*a*. The voltage across the capacitor rises at a high rate, finally causing the diode to turn off. Following that, C_L discharges through R_L. The discharge time constant, C_LR_L, is long, and the voltage falls off at a lower rate than it rose. Until time t_1 the voltage across the load follows the shape of the input signal envelope. If the discharge time constant is excessively long, the input-signal amplitude will decrease from time t_1 on (point A), and V_L will fail to follow this decrease in the envelope, thus giving rise to distortion. Between instants t_1 and t_2, the capacitor C_L discharges according to

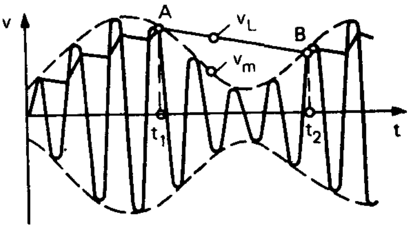


Fig. 5.11

$$V_L = V_{L1} \exp\left[-(t-t_1)/C_LR_L\right] \tag{5.30}$$

Here, V_{L1} is the load voltage at time t_1, practically equal to the input-signal amplitude, $V_m = V_{m,0}(1 + m\cos\Omega t)$.

Nonlinear distortion will not arise if C_L discharges faster than the amplitude of the input signal envelope decreases, that is, when

$$\left|\frac{dV_L}{dt}\right|_{t=t_1} \geqslant \left|\frac{dV_m}{dt}\right|_{t=t_1} \tag{5.31}$$

As follows from Eq. (5.30), V_L changes at a rate given by

$$dV_L/dt = -(V_{L1}/C_LR_L)\exp\left[-(t-t_1)/C_LR_L\right]$$

It is a maximum at the beginning ($t = t_1$) and, since $V_{L1} \approx V_{m,0} \times$ $\times\ (1 + m\cos\Omega t)$, it is equal to

$$\left|\frac{dV_L}{dt}\right|_{t=t_1} = -(V_{m,0}/C_LR_L)(1+m\cos\Omega t_1) \tag{5.32}$$

At time t_1, the signal amplitude changes at a rate given by

$$\left|\frac{dV_m}{dt}\right|_{t=t_1} = -\Omega m V_{m.0}\sin\Omega t_1 \tag{5.33}$$

On substituting Eqs. (5.32) and (5.33) in Eq. (5.31), we get the condition for freedom from distortion

$$(V_{m,0}/C_{\rm L}R_{\rm L})\,(1+m\cos\Omega t_1) \geqslant \Omega_m V_{m,0}\sin\Omega_1 t_1$$

or

$$\frac{1}{R_{\rm L}C_{\rm L}} \geqslant \frac{\Omega m\sin\Omega t_1}{1+m\cos\Omega t_1} \tag{5.34}$$

The condition defined by Eq. (5.34) must be satisfied in the worst case, which is when the right-hand side of the inequality is a maximum. In order to find the maximum, equate the derivative of the right-hand side of Eq. (5.34) with respect to time t_1 and find

$$\cos\Omega t_1 = -m$$

Then Eq. (5.34) takes the form

$$\Omega C_{\rm L}R_{\rm L} \leqslant (1-m^2)^{1/2}/m \tag{5.35}$$

The condition defined by (5.35) must be satisfied at the upper limiting frequency of modulation, $\Omega_{\rm h}$. For real signals, the modulation factor at the upper modulation frequency is seldom greater than 0.5 to 0.7. Therefore, the condition in (5.35) may be re-cast as

$$\Omega_{\rm h}C_{\rm L}R_{\rm L} \leqslant 1 \text{ to } 1.5 \tag{5.36}$$

A reduction in $R_{\rm L}$ is objectionable because this would lead to a decrease in the input resistance and detector gain. Rather, one reduces $C_{\rm L}$, but it must be 5 to 10 times the inherent diode capacitance, $C_{\rm d}$, since otherwise the voltage of the signal fed to the diode would decrease.

When the condition stated in (5.35) is satisfied, the output voltage at the upper modulation frequency faithfully reproduces the modulation, and frequency distortion is low; the condition stated in (5.35) is more rigorous than the one for freedom from distortion.

Consider the nonlinear distortion caused by the difference in load impedance to direct and alternating current. Figure 5.12 gives a family of detection curves for an ideal diode and a load line, OB, for the resistance $R_{\rm L}$ to direct current. The angle that the load line makes with the voltage axis is

$$\alpha_1 = \text{arc tan } 1/R_{\rm L}$$

With an unmodulated carrier, a direct voltage exists across the load, as determined by point A where the load line cuts the rectification curve at $V_{m.0}$. The detector output is coupled via a d.c. blocking capacitor, C_{b}, to the input of the succeeding amplifier which has a finite input resistance, R_{amp}, as shown in Fig. 5.13*a*. The value of C_{b} is chosen so as to avoid frequency distortion at the lower modulation frequencies, that is, such that

$$1/\Omega_{\mathrm{l}}C_{\mathrm{b}} \ll R_{\mathrm{amp}}$$

Therefore, the detector load resistance to the alternating current at the modulation frequency

$$R_{\Omega} = R_{\mathrm{L}}R_{\mathrm{amp}}/(R_{\mathrm{L}} + R_{\mathrm{amp}})$$

is lower than the resistance to direct current. The load line *O'B'* runs steeper (α_2 = arc tan $1/R_{\Omega}$). As the input signal amplitude

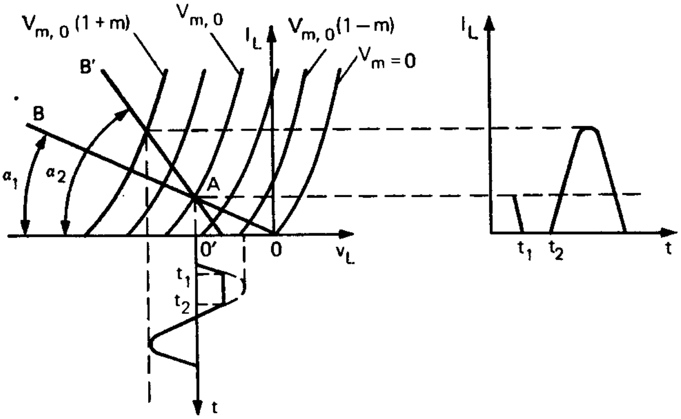


Fig. 5.12

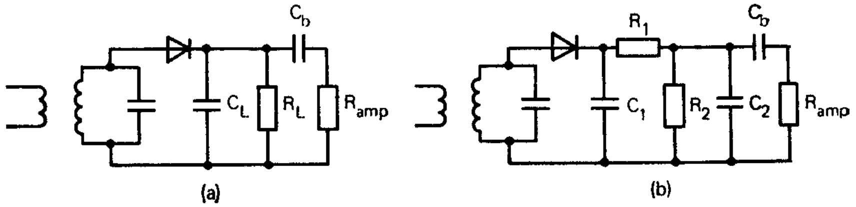


Fig. 5.13

varies from $V_{m,0}(1 - m)$ to $V_{m,0}(1 + m)$, nonlinear distortion of the envelope cutoff type will appear in the time interval from t_1 to t_2, because in the meantime the diode is non-conducting. No distortion will occur if the a.c. and d.c. load lines have the same slope, that is, when $R_{\mathrm{amp}} \gg R_{\mathrm{L}}$.

The direct component $V_m \cos\theta$ of the detector load voltage, produced by v_{in} is developed across the d.c. blocking capacitor (Fig. 5.14*a*). As the capacitor discharges, its discharge current passes through R_L and R_{amp} and produces across R_L a voltage approximately equal to $V_{m,0}$ cos $\theta R_L/(R_L + R_{amp})$. This voltage will turn off the diode if the signal amplitude falls below that value. The minimum signal amplitude is $V_{m,0}$ $(1 - m)$. Therefore, the condition for freedom from distortion due to diode cutoff may be written as

$$V_{m,0}\,(1 - m) > V_{m,0} \cos \theta R_L/(R_L + R_{amp})$$

Hence,

$$R_{amp} > R_L\,(\cos \theta + m - 1)/(1 - m)$$

For $\cos \theta \approx 1$, the above expression takes the form

$$R_{amp} > R_L m/(1 - m)$$

Notably, if m_{max} is about 0.8, then R_{amp} is greater than four times R_L. If the first stage of the amplifier that handles the detected signal is to have a sufficiently high input resistance, this stage should preferably be built around a FET. If the first stage uses a transistor, resort is made to a split-load detector, such as shown in Fig. 5.13*b*. Here, the d.c. load resistance is

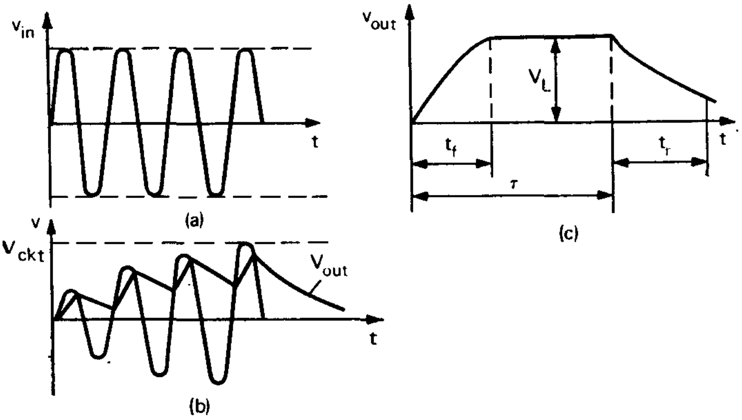


Fig. 5.14

$$R_L = R_1 + R_2$$

and the a.c. load resistance is

$$R_\Omega = R_1 + R_2 R_{amp}/(R_2 + R_{amp})$$

In such a case, it is easier to satisfy the condition for freedom from distortion, but the detector gain is reduced.

5.6. Pulse Signal Detection

There are two forms of pulse signal detection, namely:

— r.f. pulse detection in which one converts the incoming r.f. pulses to video pulses, that is, separates the envelope of each r.f. pulse in the received pulse train;

— peak detection in which one separates the envelope of the entire sequence of incoming r.f. pulses. Peak detection can be effected in two steps. Firstly, the r.f. pulses are converted to video pulses. Secondly, the video pulses are amplified and subjected to time discrimination following which the sequence of video pulses undergoes peak detection in a video amplifier.

As a rule, the pulse spacing is many times the pulse duration, therefore the detection of each r.f. pulse may be examined independently. Ordinarily it is required that, as shown in Fig. 5.14*a*, the video pulse waveform should differ only slightly from the r.f. pulse envelope. Distortion arising in pulse detection may be described in terms of the pulse rise time, t_r, and the pulse fall time, t_f, as shown in Fig. 5.14*c*. The time it takes for the load voltage to settle at its steady value depends on the rate at which the load capacitor C_L(see Fig. 5.13) discharges through the conducting diode and occupies two or three cycles of the r.f. carrier. Because the load voltage is at first low, the initial diode current cutoff angle is close to 90°, and the input resistance of the diode is low. It strongly shunts the output tuned circuit of the i.f. amplifier coupled to the detector. As the load voltage rises, the cutoff angle decreases, the input resistance of the detector increases, and the voltage across the tuned circuit, V_{ckt} tends to its steady-state value, as shown in Fig.5.14*b*.

At the instant when the incoming r.f. pulse ceases ($t = \tau$), the diode turns off, and C_L begins to discharge through R_L exponentially

$$V_{out}(t) = V_L \exp(-t/R_L C_L) \qquad (5.37)$$

The pulse fall time, t_f, is defined as that time during which the load voltage is decreasing from 90% to 10% of V_L. As follows from Eq. (5.37),

$$t_f = 2.3 R_L C_L \qquad (5.38)$$

Ordinarily, t_f is greater than t_r. Therefore, when calculating the load time constant it is usual to proceed from the allowable fall time and to invoke Eq. (5.38). The scope for the reduction of C_L is limited. The point is that if C_L be chosen lower than the diode capacitance, C_d, a sizeable proportion of the applied signal voltage would be wasted in the load. As a rule, C_{min} is taken to be equal to or greater than 5 to 10 times C_d. In consequence, one has to choose low values for R_L, and this leads to a reduction in the gain and input resistance of the detector. To avoid an excessive reduction in detector gain, the time constant R_LC_L is chosen to be greater than or equal to one or two cycles of the r.f. carrier. Then, $t_r \approx t_f$.

For a peak detector, the load time constant is chosen subject to the condition defined by Eq. (5.35). With a high pulse period-to-pulse duration ratio (that is, a low pulse duty factor), the time constant turns out to be high. For this reason, R_L is omitted, and its job is done by the reverse resistance of the diode.

5.7. Amplitude Limiters

The reception of frequency- or phase-modulated signals may be accompanied by objectionable variations in the signal amplitude. Such variations are removed by circuits known as amplitude limiters. An amplitude limiter consists of a nonlinear element and a frequency-selective network. Naturally, an amplitude limiter ought not to distort the angle modulation of the received signal. To meet this requirements, its selective network should have a bandwidth greater than the width of the signal spectrum. The performance of an amplitude limiter is evaluated in terms of its amplitude characteristic which relates the output signal amplitude to the input signal amplitude, as shown in Fig. 5.15. In an ideal amplitude limiter, any rise in the input signal amplitude over and above some threshold value, V_{th}, should leave the output signal amplitude unchanged, as represented by curve *2* in Fig. 5.15. The response of real limiters, as represented by curve *1* in the same figure, differs from the ideal one. The efficiency of a limiter is stated in terms of the ratio between the input and output modulation factors, m_{in}/m_{out}. This ratio increases in value as the suppression of the unwanted modulation becomes ever more effective.

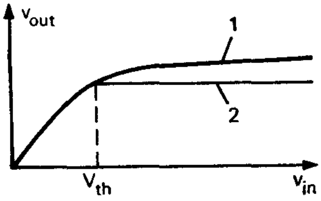


Fig. 5.15

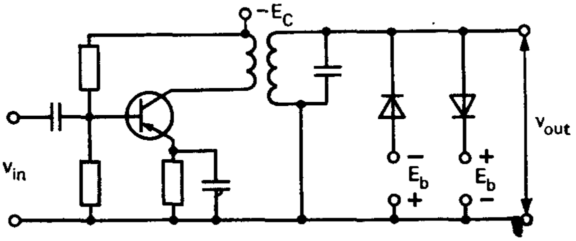


Fig. 5.16

Figure 5.16 shows the circuit of a diode limiter. It uses two diodes which are connected in parallel to the resonant circuit of a tuned amplifier and in opposition to each other, and are fed identical cutoff bias voltages, E_b. As long as the amplitude of the voltage across the resonant circuit does not exceed the cutoff bias voltage E_b, the diodes remain turned off and do not shunt the resonant circuit. Just as the signal amplitude exceeds E_b, the diodes are turned on, their input resistances shunt the resonant circuit, and the output voltage changes by a markedly smaller amount that does the input voltage.

The diode limiter has a modification known as the dynamic amplitude-modulation suppressor and shown in Fig. 5.17*a*. In contrast

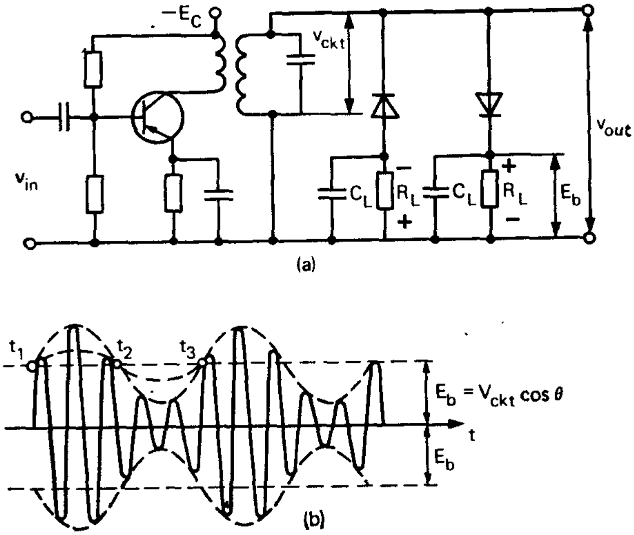


Fig. 5.17

to the circuit in Fig. 5.16, the diode circuits contain R_LC_L networks with a time constant substantially longer than it takes for the input-signal amplitude to change. As the signal is detected, a self-bias voltage is applied to the diodes, equal to

$$E_b = V_{ckt} \cos \theta$$

where V_{ckt} is the mean amplitude of the signal voltage across the tuned circuit. Because the time constant is long, the bias voltage E_b remains practically unchanged so that when the input signal amplitude increases, the cutoff angle of the diode also increases ($\cos \theta =$ $= E_b/V_{ckt}$), and the input resistance R_{in} decreases. As follows from Eq. (5.23), when $\cos \theta$ tends to zero, that is, when θ tends to $\pi/2$, the input conductance G_{in} tends to $S/2$, and the input resistance tends to $2/S$. In consequence, when the voltage across the tuned circuit increases ($V_{ckt} > E_b$), the latter is shunted increasingly more by the input resistance of the diodes (the interval between t_1 and t_2

in Fig. 5.17*b*). At low input-signal amplitudes ($V_{\text{ckt}} < E_b$), the shunting effect grows weaker (the interval between t_2 and t_3), and the output voltage only slightly varies about its mean value. The use of two diodes improves the efficiency of the limiting action.

In a transistor limiter, such as shown in Fig. 5.18*a*, limiting is effected owing to collector-current cutoff, on the one hand, and owing to transition into the saturation region, on the other, as shown in Fig. 5.18*b*. In order to bring down the limiting threshold, the transistor is operated at a reduced collector voltage. The slope of the load lines in Fig. 5.18*b* is given by

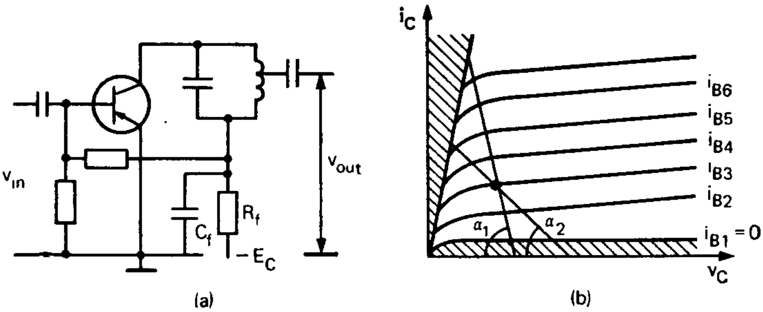


Fig. 5.18

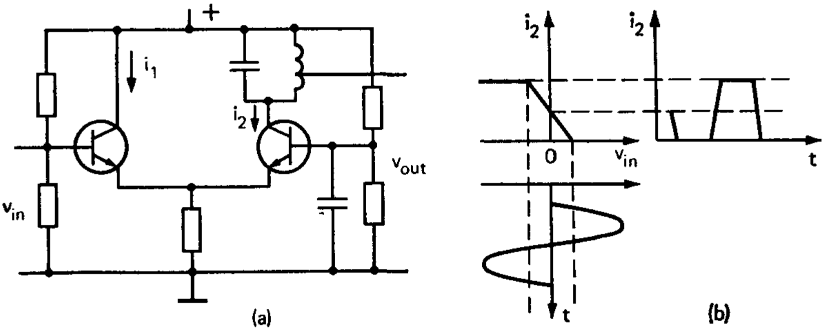


Fig. 5.19

$$\alpha_1 = \text{arc tan } (1/R_f)$$

and

$$\alpha_2 = \text{arc tan } (1/R_{\text{eq}})$$

where R_{eq} is the equivalent resonance resistance of the tuned circuit, taking into consideration all the shunting influences.

Wide use is made of limiters built around two emitter-coupled transistors, as shown in Fig. 5.19*a*, especially in IC form. The dependence of the collector current flowing in the right-hand transistor on the limiter input voltage is shown in the plot of Fig. 5.19*b*. With a high negative input voltage, the left-hand BJT is turned off and does not affect the current flowing through the right-hand transistor. As the negative input voltage decreases in magnitude, the left-hand transistor is rendered conducting, and the negative bias voltage fed from the common emitter circuit to the base of the right-hand transistor increases so that the current through the latter decreases until cutoff which occurs at a certain positive input voltage, v_{in}. When an alternating voltage is applied to the limiter input, the output current i_2 tends to a rectangular waveform as v_{in} increases. The tuned circuit in the collector lead of the right-hand transistor separates the fundamental component which remains nearly constant in magnitude as the input-signal peak exceeds the threshold value.

More efficient limiting is obtained by connecting several limiters in cascade.

5.8. Phase Detectors

A *phase detector* converts a phase-modulated voltage to a voltage varying in step with the modulating function. The output voltage

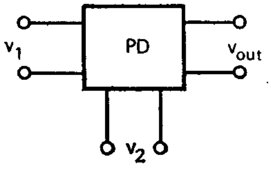


Fig. 5.20

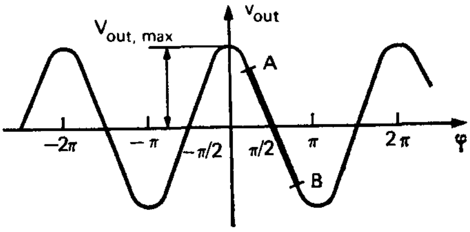


Fig. 5.21

of a phase detector is determined by the difference in phase between the two voltage waveforms being compared.

Let us draw a phase detector in the form of an equivalent six-terminal (or three-port) network, as shown in Fig. 5.20. Suppose also that it is fed two voltages

$$v_1 = V_{m1} \cos(\omega_1 t + \varphi_1)$$

and

$$v_2 = V_{m2} \cos(\omega_2 t + \varphi_2) \qquad (5.39)$$

One of them, (say, v_1) is the external received signal to be detected, and the other, v_2, is a local-source voltage used as a reference. The output voltage proportional to the phase difference is obtained by multiplying together v_1 and v_2 as follows:

$$\begin{aligned} v_{\text{out}} &= KV_{m1}V_{m2}\cos[(\omega_1 - \omega_2)t + \varphi_1 - \varphi_2] \\ &= KV_{m1}V_{m2}\cos\varphi \end{aligned} \tag{5.40}$$

Thus, a phase detector is, in effect, a voltage multiplier. In Eq. (5.40) K is a proportionality factor and φ is the instantaneous phase difference between the voltages being compared. It may be resolved into two components

$$\varphi_\omega = (\omega_1 - \omega_2)t$$

and

$$\varphi_0 = \varphi_1 - \varphi_2$$

The former is due to the difference in frequency between v_1 and v_2; the latter is equal to the difference between their epochs. In the detection of phase-modulated signals, it is essential to arrange so that $\omega_1 = \omega_2$. If one of the two voltages is first shifted in phase through $\pi/2$, then

$$v_{\text{out}} = KV_{m1}V_{m0}\sin\varphi$$

At low values of ω, it may be taken that

$$v_{\text{out}} \approx KV_{m1}V_{m2}\varphi$$

or that the output voltage is a faithful replica of the modulating function.

The performance of a phase detector is basically described by its detection characteristic which relates the output voltage v_{out} to the phase difference φ between the signal and reference voltages. The detection characteristic of an ideal phase detector (or voltage multiplier) is described by Eq. (5.40) and has the shape shown in Fig. 5.21.

The key parameters of a phase detector are the slope of the detection characteristic, the voltage gain, and distortion in the detection of continuous (analog) signals.

The slope of the characteristic is the output voltage derivative with respect to the phase angle at a maximum of the derivative with specified input-signal amplitudes

$$S_{\text{PD}} = \left|\frac{dV_{\text{out}}}{d\varphi}\right|_{\text{max}}$$

The voltage gain of a phase detector is given by

$$K_{\text{PD}} = V_{\text{out,max}}/V_{m1}$$

Distortion in the detection of analog signals depends on the linearity of the working portion of the detection characteristic (say, portion AB in Fig. 5.21).

5.9. Phase Detector Types

Wide use is made of *balanced phase detectors* arranged as shown in Fig. 5.22. A balanced phase detector is a pair of unbalanced detectors interconnected so that their output voltages give the difference between the detected signals. Its inputs are fed the voltages defined by Eq. (5.39). The external signal, v_1, is applied in anti-phase, and the local-source or reference voltage, v_2, is applied in phase. The amplitudes of the voltages applied to the diodes can be found from the phasor diagram in Fig. 5.23:

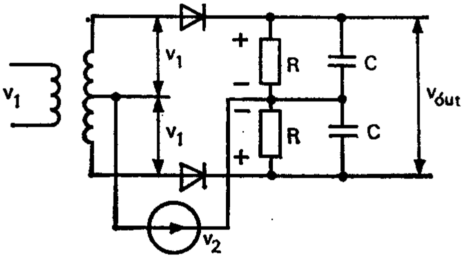


Fig. 5.22

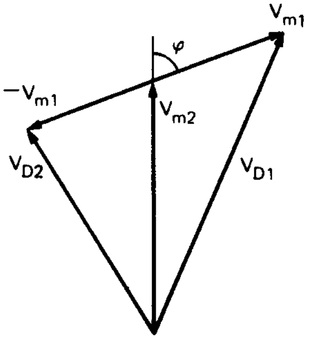


Fig. 5.23

$$\left.\begin{aligned} V_{D1}&=(V_{m1}^2+V_{m2}^2+2V_{m1}V_{m2}\cos\varphi)^{1/2} \\ V_{D2}&=(V_{m1}^2+V_{m2}^2-2V_{m1}V_{m2}\cos\varphi)^{1/2} \end{aligned}\right\} \qquad (5.41)$$

The two voltages, V_{D1} and V_{D2}, are detected and produce across the respective loads two output voltages

$$V_{out1} = K_d V_{D1}$$

and

$$V_{out2} = K_d V_{D2}$$

where K_d is the voltage gain of the amplitude detectors. In agreement with Eq. (5.41), the resultant voltage is

$$\begin{aligned} V_{out} &= (V_{D1} - V_{D2})\, K_d \\ &= K_d\, [(V_{m1}^2 + V_{m2}^2 + 2V_{m1}V_{m2}\cos\varphi)^{1/2} \\ &\quad - (V_{m1}^2 + V_{m2}^2 - 2V_{m1}V_{m2}\cos\varphi)^{1/2}] \end{aligned} \qquad (5.42)$$

The shape of the characteristic depends on the relative magnitude of V_{m1} and V_{m2}. The dependence of v_{out} on φ in the range from 0 to π is most linear when V_{m1} and V_{m2} are equal, as shown in Fig. 5.24.

If $V_{m2} \gg V_{m1}$, then, putting V_{m2} outside the brackets and neglecting the term $(V_{m1}/V_{m2})^2 \ll 1$, we obtain

$$v_{\text{out}} \approx K_d V_{m2} \{[1 + 2 (V_{m1}/V_{m2}) \cos \varphi]^{1/2} - [1 - 2 (V_{m1}/V_{m2}) \cos \varphi]^{1/2}\} \quad (5.43)$$

Now we expand each term in the square brackets by the binomial theorem and limit ourselves to the first two terms of each series. After simple manipulations, the expression for the detection characteristic takes the form

$$v_{\text{out}} = 2K_d V_{m1} \cos \varphi \quad (5.44)$$

As is seen, when $V_{m1} \ll V_{m2}$, the detection characteristic is a cosinusoid very nearly (see Fig. 5.21), varies linearly with the amplitude

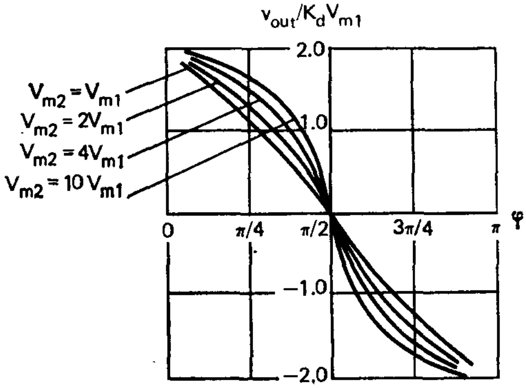


Fig. 5.24

of the lower (external-signal) voltage, and is independent of the amplitude of the higher (local-source or reference) voltage.

Since the output voltage of a phase detector varies linearly with its input voltage, it can be used to detect amplitude-modulated signals; that is how the *synchronous detector* is implemented (see Sec. 5.1). The reference voltage supplied by a local source must be synchronized with the input-signal carrier accurate to within the phase. The output voltage of the detector is maximal when the phase difference is zero, $\varphi = 0$. When $\varphi = 90°$, there is no output voltage, whereas at $\varphi = 180°$, the polarity of output voltage is reversed. Since the detector is symmetrical relative to the applied voltages, it is immaterial which voltage is applied to which input.

In some cases, it is required that the phase detector should filter out combination frequencies other than $\omega_1 - \omega_2$. This is where resort is made to the *ring phase detector* shown in Fig. 5.25. A ring phase detector may be looked upon as a combination of two balanced detectors operating into a common load. One detector is formed by diodes D_1 and D_2, and the other by diodes D_3 and D_4. With all other conditions being equal, the output voltage of a ring detector is nearly half as high, but the diodes placed in the diagonals cancel out the even harmonics of the input signals.

The gain and input resistance of a phase detector can be increased if the diodes are replaced by amplifying devices. Figure 5.26 shows the circuit of a balanced phase detector which uses FETs in the switching mode. The input voltage v_1 is applied to the FET gates in anti-phase, while the local-source reference voltage v_2 is applied to the FET drains in phase. It must be high enough to render the respective FET conducting during one of its half-cycles. The output voltage varies with the phase shift between v_1 and v_2 in the same way as in the circuit of Fig. 5.21.

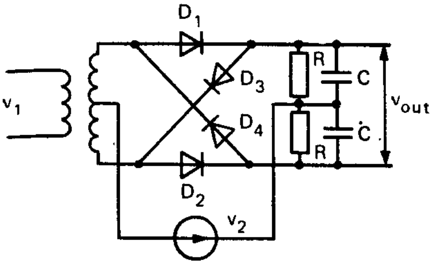


Fig. 5.25

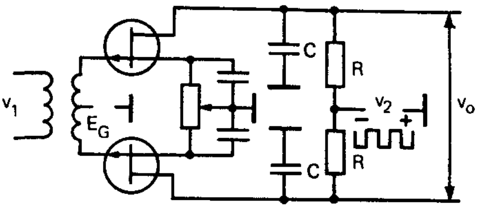


Fig. 5.26

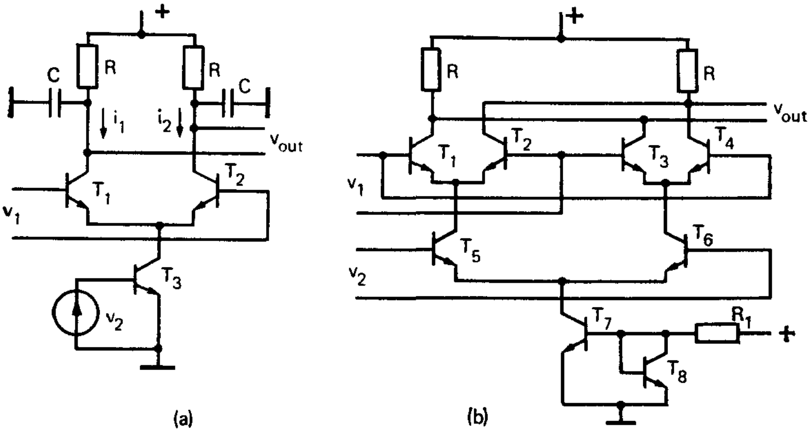


Fig. 5.27

Integrated-circuit (IC) detectors, widely used as multipliers, depend for their operation on the control of the transconductance of a differential transistor pair arranged as shown in Fig. 5.27*a*. This type of detector is not unlike the balanced transistor frequency converter shown in Fig. 4.15. The difference is that instead of a filter tuned to the intermediate frequency, the phase detector is loaded into *RC* networks which act as low-pass filters. The detector is fed the external received-signal voltage

$$v_1 = v_C = V_{m,s} \cos (\omega_s t + \varphi_s) \tag{5.45}$$

and a reference voltage from the local source

$$v_2 = v_r = V_r \cos \omega_r t$$

The signal voltage is applied to the bases of transistors T_1 and T_2 in anti-phase, whereas the reference voltage is applied in phase, thus causing identical changes in their transconductance, S*. Therefore, the intermodulation components of currents i_1 and i_2 are in anti-phase, that is,

$$i_1 = -i_2 = Sv_1$$

or, subject to Eq. (5.45) (see also Sec. 4.7),

$$i_1 = -i_2 = \left(S_0 + \sum_{k=1}^{\infty} S_{mk} \cos k\omega_r t\right) V_{m,s} \cos (\omega_s t + \varphi_s)$$

The output voltage is produced by the difference between the direct components of currents i_1 and i_2, that is,

$$v_{out} = S_{m1} V_{m,s} R \cos \varphi \tag{5.46}$$

where

$$\varphi = (\omega_s - \omega_r) t + \varphi_s$$

Equation (5.46) is analogous to Eq. (5.44).

The multiplier shown in Fig. 5.27*a* spans a small dynamic range of input-signal levels and operates only within two quadrants of its characteristic. In practice, wider use is made of the *double balanced multiplier* shown in Fig. 5.27*b* which uses three differential transistor pairs. In fact, it is a combination of two balanced circuits operating into common loads, *R*. The signal voltage v_1 is applied to transistor pairs T_1/T_2 and T_3/T_4 whose transconductances are varied under the action of the reference voltage v_2 applied to the bases of transistors T_5 and T_6. In each transistor pair, the signal voltage is applied in anti-phase, and the reference voltage is applied in phase to the two transistors of one pair and in anti-phase to the transistors in the different pairs. The transistor currents are determined by a d.c. generator built around transistor T_7 whose base voltage is stabilized by a network made up of resistor R_1 and diode-connected transistor T_8. A major advantage of this type of multiplier is the

* As has been noted earlier, many texts designate the transconductance as g_m.— *Translator's note.*

fact that it can effect multiplication in all the four quadrants of its characteristic. Practical multipliers include stages to effect transition from an unbalanced to a balanced connection and back. The output voltage is smoothened by a low-pass filter.

When a phase detector operates as a switch, if one of the applied voltages controls the switching action, the output voltage is independent of this control voltage. If both (signal and reference) voltages solely serve to switch the amplifying devices, the output voltage is independent of both voltages. Figure 5.28 shows the circuit of a switching detector built around three differential transistor pairs. Transistor T_1 supplies a stable (regulated) direct current, I. Its base voltage is stabilized by a network consisting of a resistor R_1 and a diode-connected transistor, D_1. The current I passes through T_2 and T_3. In turn, the currents I_1 and I_2 of these transistors flow, respectively, through transistors T_4 and T_5 (I_1' and I_1''), and transistors T_6 and T_7 (I_2' and I_2''). Transistors T_2 through T_7' operate as switches controlled by the reference and signal voltages, v_r and v_s. For this purpose, limiters *1* and *2* shape v_r and v_s into rectangular pulses. The waveforms of currents and voltages are shown in Fig.5.29.

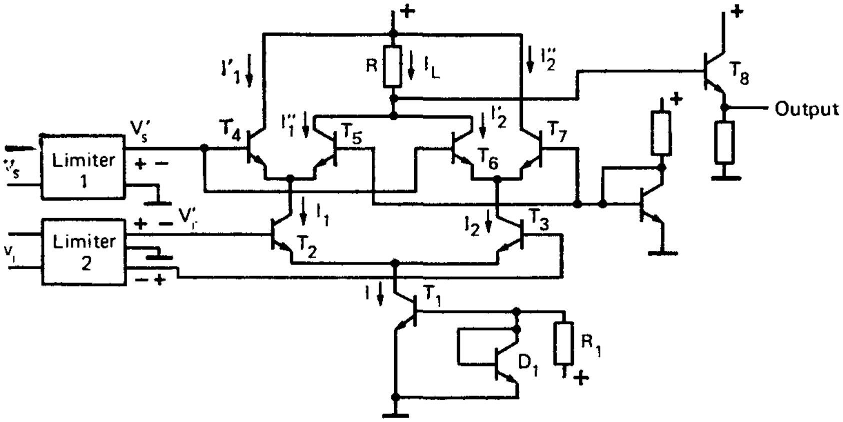


Fig. 5.28

When v_r is positive, the current I flows through transistor T_2; when v_r is negative, the current flows through transistor T_3. When v_s is positive, the currents I_1 and I_2 pass through transistors T_4 and T_5; when v_s is negative, they flow through transistors T_5 and T_7. The currents I_1 and I_2 are pulses with a duration of a half-cycle. When there is a phase shift, φ, between the signal voltage v_s and the reference voltage v_r, currents flow through transistors T_4 through T_7 as follows: current I_1' flows through T_4 when v_s' and v_r' are positive; current I_1'' flows through T_5 when v_s' is negative and v_r' is negative; current I'_2 flows through T_6 when v'_s is positive and v'_r is negative; current I''_2 flows through T_7 when both v'_s and v'_r are negative. Pulses of currents I'_1 and I''_2 have a width (or duration) equal to $\pi - \varphi$, whereas pulses of currents I''_1 and I'_2 have a width equal to φ. The load resistor carries a sum current

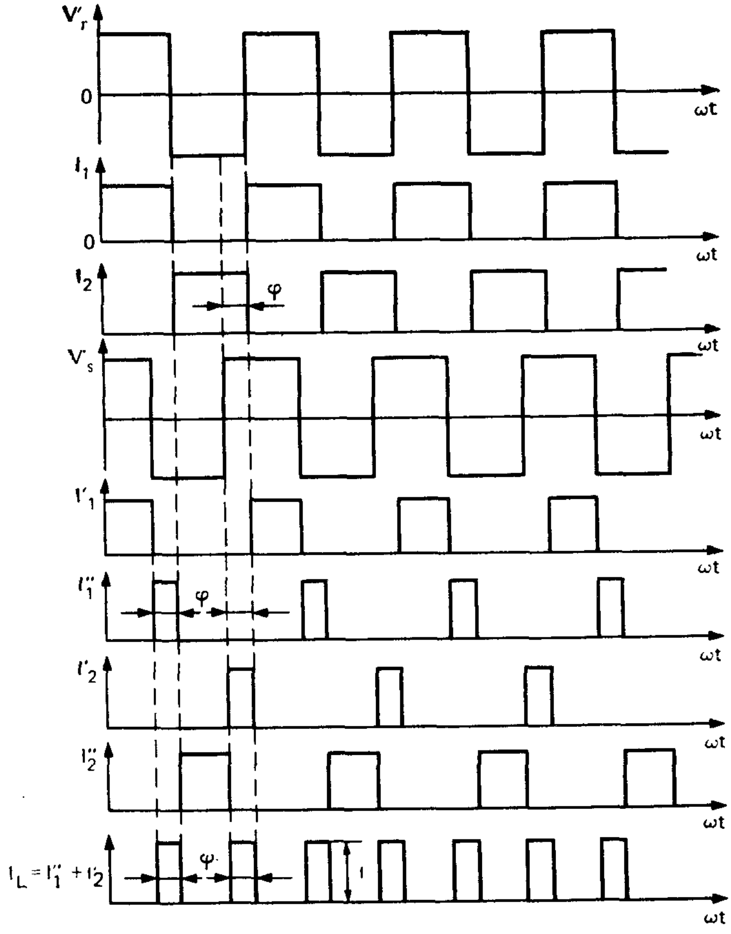


Fig. 5.29

$$I_L = I''_1 + I'_2$$

constituted by a train of pulses of width φ, recurring every half-cycle. The mean value of this current is proportional to φ. The characteristic of this detector is shown in Fig. 5.30. The output voltage is picked off the emitter follower built around T_8 and smoothened by a low-pass filter.

5.10. Principles of Frequency Detection

According to the principle of operation, there may be amplitude frequency detectors, phase frequency detectors, and pulse frequency detectors. In an amplitude frequency detector, variations in the signal frequency are converted to amplitude variations, and these are then subjected to amplitude detection. In a phase frequency detector, frequency variation are converted to variations in the phase shift between two voltages, and these variations are then subjected to phase detection. In a pulse frequency detector, a frequency-modulated (FM) signal is converted to a sequence of pulses with a repetition rate proportional to the deviation of the input signal frequency from the centre frequency. The output voltage proportional to the number of pulses per unit time can be derived with the aid of a pulse counter. For this reason, such devices are called *pulse-counting detectors*.

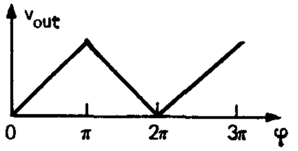


Fig. 5.30

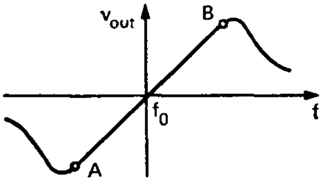


Fig. 5.31

The characteristic of a frequency detector relates output voltage to .signal frequency, with the amplitude of input voltage held constant, as shown in Fig. 5.31. Detection quality is determined by the linearity of the operating portion, *AB*, of the characteristic. An important parameter of a frequency detector (or discriminator) is the slope of its characteristic

$$S_{FD} = \left|\frac{dv_{out}}{df}\right|_{f=f_0}$$

5.11. Types of Frequency Detectors (Discriminators)

Figure 5.32 shows the circuit of a widely used frequency detector known as the *double-tuned frequency discriminator*. In this arrangement, two tuned circuits are used, one tuned to a frequency slightly above the centre frequency of the received signal

$$f_1 = f_0 + \Delta f_0$$

and the other to a frequency slightly below the centre frequency of the signal

$$f_2 = f_0 - \Delta f_0$$

As the signal frequency rises, it moves closer to the resonant frequency of the first tuned circuit, f_1, and away from the resonant frequency of the second, f_2. The voltage across the first tuned circuit increases, and that across the second decreases. As the signal frequency decreases, it moves closer to f_2 and away from f_1. In this way, the FM signal becomes amplitude- and frequency-modulated. The voltages appearing across the tuned circuits are applied to amplitude diode detectors. The resultant voltage is the difference of two tuned-circuit voltages

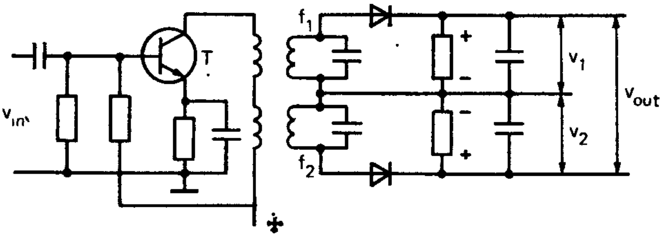


Fig. 5.32

$$v_{\text{out}} = v_1 - v_2 = K_{\text{d}}\,(V_{\text{ckt1}} - V_{\text{ckt2}}) \tag{5.47}$$

where K_{d} is the gain of the diode detectors.

In Fig. 5.33, the voltages across the diode detector loads, v_1 and v_2, are shown in an appropriate polarity by the dashed curves, and the output voltage, by the full curve. In Eq. (5.47), the voltages across the first and second tuned circuits, are, respectively, given by

$$\begin{aligned} V_{\text{ckt1}} &= V_0/[1 + (2\Delta f_1/fd_{\text{eq1}})^2]^{1/2} \\ &= V_0/[1 + (\xi_{01} - \xi_1)^2]^{1/2} \end{aligned} \tag{5.48}$$

$$\begin{aligned} V_{\text{ckt2}} &= V_0\,[1 + (2\Delta f_2/f_2 d_{\text{eq2}})^2]^{1/2} \\ &= V_0/[1 + (\xi_{02} + \xi_2)^2]^{1/2} \end{aligned} \tag{5.49}$$

where $\Delta f_1 = \Delta f_0 - \Delta f$, $\Delta f_2 = \Delta f_0 + \Delta f$ are the absolute amounts off resonance, detunings, or *frequency offsets* for the two tuned circuits with the signal frequency deviation equal to Δf, and

$$V_0 = mV_{\text{in}}\mid y_{21}\mid R_{\text{eq}} \tag{5.50}$$

is the amplitude of the voltage across each tuned circuit at resonance. If the transistor T_1 operates in the cutoff region, $\mid y_{21}\mid$ in Eq. (5.50) should be replaced by S_{m1}, or the slope at the fundamental frequency.

In Eqs. (5.48) and (5.49), $\xi_{01} = 2\Delta f_0/f_1 d_{\text{eq1}}$ and $\xi_{02} = 2\Delta f_0/f_2 d_{\text{eq2}}$ are generalized frequency offsets. For the voltage-vs-frequency characteristic of the frequency discriminator to be symmetrical, it is essential that

$$\xi_{01} = \xi_{02} = \xi_0$$

that is,

$$2\Delta f_0/f_1 d_{eq1} = 2\Delta f_0/f_2 d_{eq2}$$

In other words, the two tuned circuits must have the same bandwidth

$$f_1 d_{eq1} = f_2 d_{eq2} = f_0 d_{eq}$$

Therefore, with any frequency deviation, Δf, the generalized frequency offset or detuning is

$$\xi_1 = \xi_2 = \xi = 2\Delta f/f_0 d_{eq}$$

In view of the foregoing and on substituting Eqs. (5.48) and (5.49)

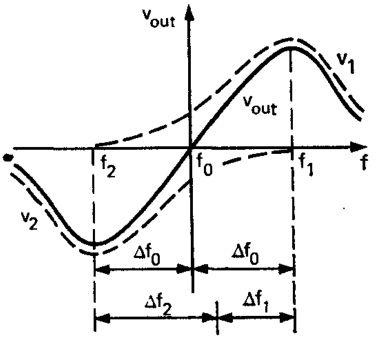


Fig. 5.33

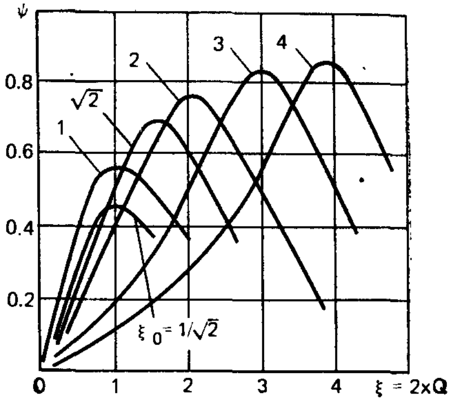


Fig. 5.34

in Eq. (5.47), we obtain the following expression for the output voltage of the double-tuned frequency discriminator

$$v_{out} = mV_{in} \mid y_{21} \mid R_{eq} K_d \psi(\xi) \qquad (5.51)$$

where

$$\psi(\xi) = 1/[1 + (\xi_0 - \xi)^2]^{1/2} - 1/[1 + (\xi_0 + \xi)^2]^{1/2} \qquad (5.52)$$

is a functiod of generalized detuning or a normalized voltage-vs-frequency characteristic of the discriminator. Figure 5.34 shows the right-hand half of the $\psi(\xi)$ characteristic for several values of ξ_0. Since the characteristic is symmetrical, nonlinear distortion can arise solely due to the odd harmonics of the modulating frequency. The even harmonics are felt only when the two tuned circuits of the discriminator are not identical or when the discriminator is not properly tuned. The voltage-vs-frequency characteristic of the double-tuned frequency discriminator comes closest to a linear one when $\xi_0 \approx \sqrt{1.5}$.

A popular form of frequency discriminator, known as the *Foster-Seeley discriminator* and shown in Fig. 5.35, uses two coupled tuned circuits L_1C_1 and L_2C_2. This arrangement falls in the category of phase-frequency detectors. Demodulation is effected by the L_2C_2 circuit tuned to the centre frequency of the received signal. When

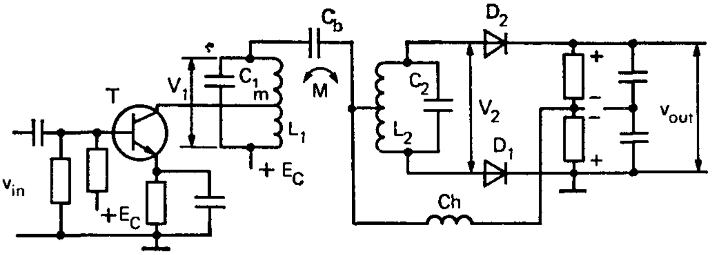
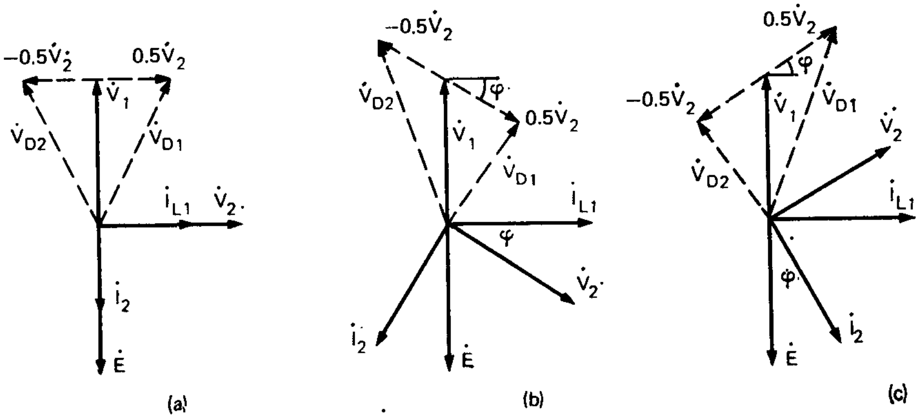


Fig. 5.35



(a) (b) (c)

Fig. 5.36

an unmodulated carrier arrives at the input of the discriminator, the voltage across the second tuned circuit, $\dot{V}_2$, is shifted through 90° relative to that across the first, $\dot{V}_1$. When an FM signal is received, an additional phase shift, φ, proportional to the frequency deviation appears between $\dot{V}_1$ and $\dot{V}_2$. Let us prove the point by reference to the phasor diagrams in Fig. 5.36. Let the phasor $\dot{V}_1$ be the reference, or datum, phasor (see Fig. 5.36*a*). The current I_{L1} in the coil L_1 is in quadrature lagging with $\dot{V}_1$. In the second tuned circuit this current induces an emf equal to

$$\dot{E} = -j\omega M\dot{I}_{L1}$$

which gives rise to a current, $\dot{I}_2$. At resonance, $\dot{I}_2$ is in phase with $\dot{E}$ and produces across L_2 a voltage, $\dot{V}_2$, which is in quadrature leading with respect to $\dot{I}_2$. That is why $\dot{V}_1$ and $\dot{V}_2$ are 90 degrees out of phase with each other at resonance.

If the signal frequency is higher than the resonant frequency of the tuned circuits, $f_s > f_0$, the current $\dot{I}_2$ now lags behind $\dot{E}$ in phase by some angle φ, as shown in Fig. 5.36*b*, because the impedance of the second tuned circuit is inductive in its effect. The voltage $\dot{V}_2$ is still in quadrature leading with $\dot{I}_2$, so $\dot{V}_2$ is shifted relative to $\dot{V}_1$ by an angle exceeding 90°. Similarly it may be shown that when $f_s < f_0$, the phase shift between $\dot{V}_1$ and $\dot{V}_2$ is less than 90°, as is seen from Fig. 5.36*c*. Thus, frequency variations are converted to variations in the phase shift between the voltages across the first and second tuned circuits. These voltages are applied to the diodes of a phase detector arranged as shown in Fig. 5.22. Here, $\dot{V}_1$ is used as reference and applied to the diodes in phase. It is picked off the first tuned circuit via a d.c. blocking capacitor, C_b. The voltage $\dot{V}_2$ is applied to the diodes in antiphase. Across each diode, the voltage is the vectorial sum of the first tuned-circuit voltage and a half of the second tuned-circuit voltage

$$\dot{V}_{d1} = \dot{V}_1 + 0.5\dot{V}_2$$

and

$$\dot{V}_{d2} = \dot{V}_1 - 0.5\dot{V}_2$$

as shown in Fig. 5.36*c*. The output voltage is the difference between the rectified voltages multiplied by the detector gain

$$V_{out} = (|\dot{V}_{d1}| - |\dot{V}_{d2}|)\, K_d$$

When the received signal is at the centre frequency of the secondary (or, which is the same, when an unmodulated carrier is applied to the discriminator input), the output voltage, V_{out}, is zero, because each diode is connected across one half of the secondary and the primary in series, the resultant r.f. voltages developed across each diode load resistor are equal and of opposite polarity, $|\dot{V}_{d1}| = |\dot{V}_{d2}|$, and the net voltage between the top of the load resistors and ground is zero. This is shown vectorially in Fig. 5.36*a*. If, however, the signal frequency varies from the resonant frequency (or, which is the same, when the carrier is modulated), there is a change in the phase relationship between $\dot{V}_1$ and $\dot{V}_2$ and an accompanying change in $\dot{V}_{d1}$ and $\dot{V}_{d2}$, as shown vectorially in Fig. 5.36*b* and *c*.

The value and polarity of the output voltage depend on the magnitude and direction of change in the signal frequency. The voltage-vs-frequency characteristic of this discriminator is similar to that shown in Fig. 5.31.

For purposes of analysis, let us draw an equivalent circuit for the Foster-Seeley discriminator as shown in Fig. 5.37. Here, transistor T_1 is replaced by an equivalent current generator $\dot{Y}_{21}\dot{V}_{\text{in}}$ of equivalent admittance

$$\dot{Y}_{22} = G_{22} + j\,\omega C_{22}$$

Capacitance C_{22} is taken care of by tuning the first tuned circuit as appropriate, and the conductance G_{22} is related to the equivalent damping factor of the tuned circuit. Let us reference the equivalent current generator parameters to all of the tuned circuit and draw an equivalent voltage generator circuit as shown in Fig. 5.38. Here, $\dot{E}_1$ is found in the same manner as in Sec. 3.11 (see Fig. 3.27), namely:

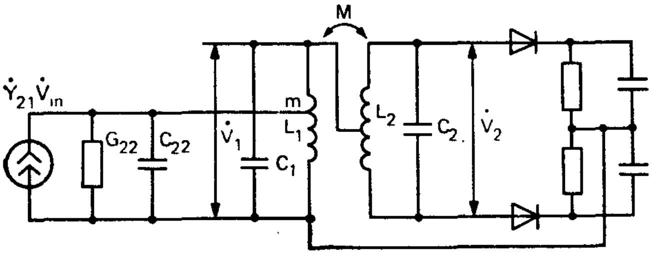


Fig. 5.37

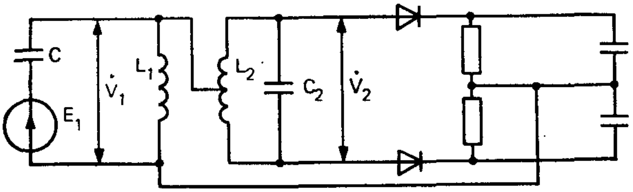


Fig. 5.38

$$\dot{E}_1 = m\dot{Y}_{21}\dot{V}_{\text{in}}/j\,\omega C$$

where

$$C = C_1 + m^2 C_{22} + C_{\text{w}}$$

where C_w is the wiring capacitance. Let us write Kirchhoff's equations for coupled tuned circuits

$$\dot{E}_1 = \dot{I}_1\dot{Z}_1 - \dot{I}_2 j\,\omega M$$

$$0 = \dot{I}_2\dot{Z}_2 - \dot{I}_1 j\,\omega M$$

and solve them for the currents

$$\left.\begin{aligned}\dot I_1 &= \dot E_1\dot Z_2/(\dot Z_1\dot Z_2+\omega^2M^2)\\ \dot I_2 &= \dot E_1 j\omega M/(\dot Z_1\dot Z_2+\omega^2M^2)\end{aligned}\right\}\qquad(5.53)$$

where

$$\left.\begin{aligned}\dot Z_1 &= \rho_1(d_{\text{eq}1}+jy)\\ \dot Z_2 &= \rho_2(d_{\text{eq}2}+jy)\end{aligned}\right\}\qquad(5.54)$$

are the tuned-circuit impedances.

The voltages across the diodes are found to be

$$\begin{aligned}\dot V_{\text{d}1} &= \dot I_1 j\omega L_1 + 0.5\dot I_2 j\omega L_2\\ \dot V_{\text{d}2} &= \dot I_1 j\omega L_1 - 0.5\dot I_2 j\omega L_2\end{aligned}\qquad(5.55)$$

Assume that the two tuned circuits have identical parameters ($C_1 \approx C_2 = C$, $L_1 \approx L_2 = L$, $d_{\text{eq}1} \approx d_{\text{eq}2} = d_{\text{eq}}$), as is usually done in practice. Also, near resonance $\omega_0/\omega \approx 1$. On substituting Eqs. (5.53) and (5.54) in Eq. (5.55) and noting the expression for $\dot E_1$, we obtain

$$\begin{aligned}\dot V_{\text{d}1} &= m\dot Y_{21}\dot V_{\text{in}}\rho\,\frac{d+j(y+0.5k)}{(d+jy)^2+k^2}\\ \dot V_{\text{d}2} &= m\dot Y_{21}\dot V_{\text{in}}\rho\,\frac{d+j(y-0.5k)}{(d+jy)^2+k^2}\end{aligned}$$

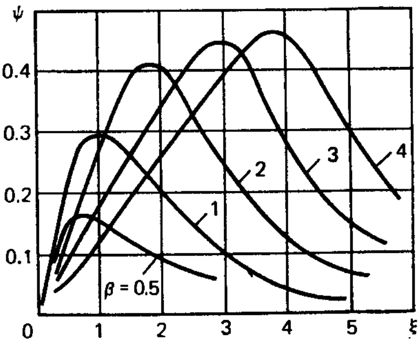


Fig. 5.39

Taking the magnitudes of $\dot V_{\text{d}1}$ and $\dot V_{\text{d}2}$ and carrying out simple manipulations, we find the output voltage to be equal to

$$\dot v_{\text{out}} = K_{\text{d}}(|\dot V_{\text{d}1}| - |\dot V_{\text{d}2}|) = m\dot Y_{21}R_{\text{eq}}V_{\text{in}}K_{\text{d}}\psi(\xi)\qquad(5.56)$$

where

$$\psi(\xi) = \frac{[1+(\xi+0.5\beta)^2]^{1/2}-[1+(\xi-0.5\beta)^2]^{1/2}}{[(1+\beta^2-\xi^2)^2+4\xi^2]^{1/2}}\qquad(5.57)$$

Here, $\xi = y/d$ and $\beta = k/d$.

The function $\psi(\xi)$ is the normalized characteristic of the discriminator. It is symmetrical about the origin of coordinates. The right-hand part of a family of $\psi(\xi)$ curves is shown in Fig. 5.39. The detector characteristic is most linear within its middle portion at $\beta = 0.5$ to 2.

The frequency detectors (discriminators) we have just examined require that the applied signal should first be amplitude-limited. The Foster-Seeley discriminator shown in Fig. 5.35 may be re-arranged so that it will also double as a limiter. For this purpose, one of the diodes is connected in reversed polarity, the load resistors R_1 and R_2 are shunted by a high-value capacitor C_0, and the detected voltage is picked off between the junctions of C_1/C_2 and R_1/R_2, as shown in Fig. 5.40. The voltages across the diodes answer the phasor diagrams in Fig. 5.36 as before. The direct component of current of the two diodes flows in a common circuit containing D_1, L_2, D_2, R_1, and R_2, and produces a voltage drop across R_1 and R_2. Since the $R_1/R_2/C_0$ network has a long time constant, the voltage V_0 is maintained equal to the mean signal voltage for a relatively long time. Therefore, variations in the input signal amplitude cause changes in the current cutoff angles defined by Eq. (5.14) and, in consequence, changes in the input resistances of the diode detectors. Amplitude modulation is suppressed in the same way as in the limiter of Fig. 5.17 owing to the fact that the detector tuned-circuits are shunted to a varying degree as the signal amplitude varies. For example, when the signal amplitude increases, cos θ decreases, the cutoff angle of each diode increases, the input resistances of the diode detectors go down and shunt the discriminator tuned circuits progressively more. When the signal amplitude decreases, the input resistances of the diode detectors shunt the tuned circuits to a lesser degree. To avoid any overcompensation of amplitude variations, which is likely to occur in this case, it is usual to include low-value resistors (shown dashed in Fig. 5.40) which are unbypassed by the high-value capacitor C_0.

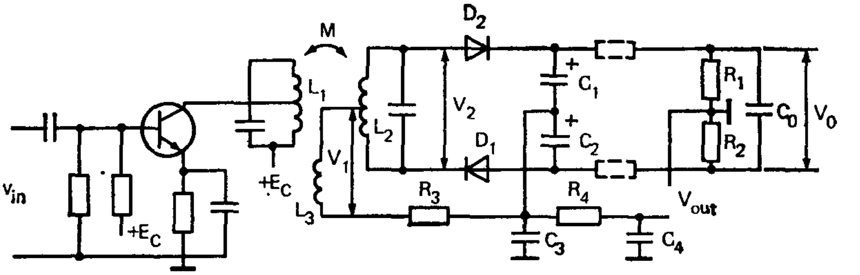


Fig. 5.40

The alternating components of diode currents flow in the circuits consisting of D_1, L_2, L_3, R_3, C_2 and D_2, C_1, R_3, L_3, L_2, respectively, and produce across C_1 and C_2 two audio-frequency voltages, v_{C1} and v_{C2}. With frequency modulation, although their sum remains constant, $v_{C1} + v_{C2} = \text{const}$, the ratio of the two audio output voltages varies. Hence the name '*ratio detector*'. The output voltage s picked off between the junctions of C_1/C_2 and R_1/R_2 and is equal to

$$v_{\text{out}} = V_{\text{C1}} - v_{\text{R1}} = v_{\text{C1}} - 0.5V_0$$

Since

$$V_0 = v_{\text{C1}} + v_{\text{C2}}$$

the output voltage of a ratio detector

$$v_{\text{out}} = 0.5\,(v_{\text{C1}} - v_{\text{C2}})$$

is half as great as that of the Foster-Seeley discriminator. The output voltage of a ratio detector goes to an a.f. amplifier via a de-emphasis network, R_4C_4, which removes the pre-emphasis applied at the

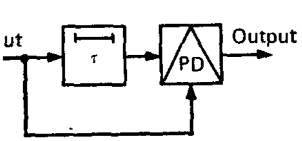


Fig. 5.41

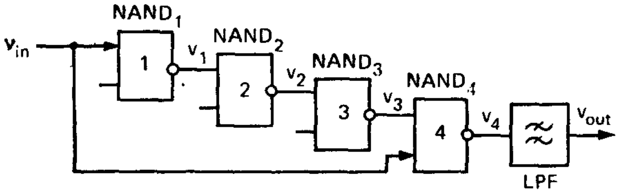


Fig. 5.42

transmitter in order to emphasize the higher modulating frequencies. In more detail pre-emphasis will be discussed in Sec. 9.4. The purpose of resistor R_3 in the circuit of Fig. 5.40 is to avoid untoward resonances in the L_3C_3 network.

Combining the functions of a limiter and a detector in a single arrangement makes it less expensive, but the quality of limiting and detection may be better if the two functions are separated. For this reason, the ratio detector is mostly used in inexpensive broadcast receivers.

Figure 5.41 shows a frequency detector now available in a single IC package. In this arrangement, frequency modulation is converted to phase modulation by a delay element which produces a phase shift, $\varphi = \omega\tau$, proportional to frequency, rather than by a tuned circuit. The required reference voltage is provided by the input signal. The output voltage of a phase detector is a function of the phase shift and is, therefore, proportional to frequency over a certain range. Since after it has been clipped, the signal current reaching the input of the frequency detector has the shape of a nearly rectangular pulse, it is a simple matter to provide the required delay by means of binary logic gates.

Figure 5.42 shows a frequency detector which uses a 4-NAND IC. The delay is provided by three NAND gates in which the second inputs are left floating, so that they are set to a logic *1* potential. The fourth NAND gate operates as a switching phase detector (a coincidence stage). The voltage waveforms existing in the various portions of the circuit for the two input-signal frequencies ($f_1 > f_2$) are shown in Fig. 5.43. During the positive half-cycles of the input signal with an amplitude exceeding the threshold voltage of the first logic gate, its output is set to a logic 0. The output signal of the gate is shifted from the input one by time τ. On passing through three logic gates with a total delay of 3τ, the signal arrives at the input to the fourth

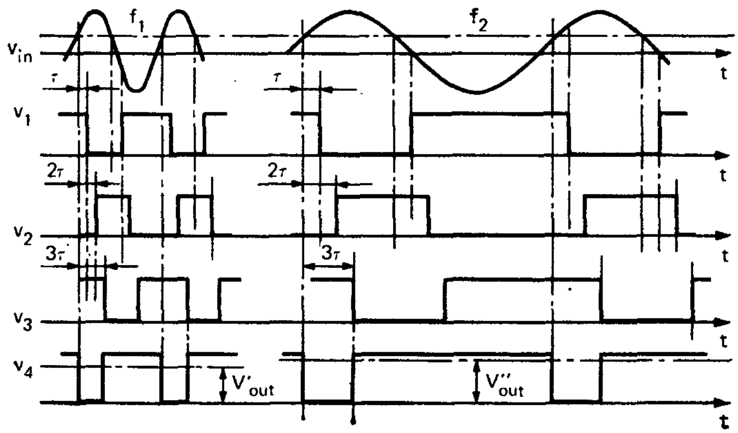


Fig. 5.43

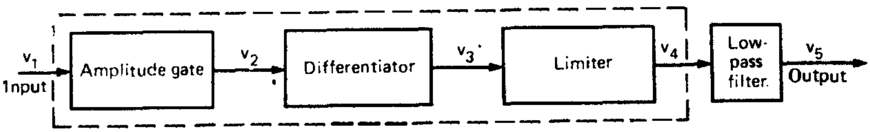


Fig. 5.44

logic gate which, as has been noted, operates as a coincidence stage· This stage generates pulses with a duration inversely proportional to the signal frequency. There is a low-pass filter which separates the mean pulse voltage. The slope of the detection characteristic is proportional to the delay time. The relation between the output voltage and frequency would cease to be linear if the resultant delay time $3\tau_0$ were in excess of a half of the input-signal period. To avoid this, it is usual to take $f_{\max} \leqslant 1/6\tau_0$. This type of detector has found use in TV receivers, multichannel radio-relay links and satellite communication systems.

In conclusion, let us examine in brief *pulse-counting detectors.* In a pulse-counting detector, an FM signal is converted to a train of pulses of constant amplitude and duration. The pulse repetition frequency is made to vary with the frequency of the input signal which means that an FM signal is converted to a pulse-position-modulated (PPM) signal. The sequence of pulses is averaged to yield a voltage proportional to the pulse repetition rate. In block-diagram

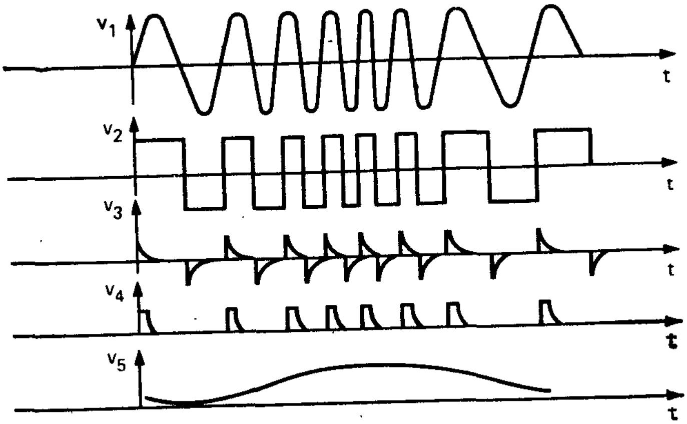


Fig. 5.45

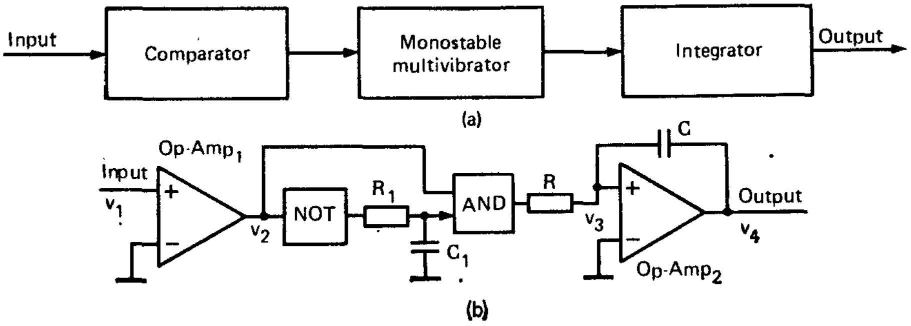


Fig. 5.46

form, a pulse-counting detector is shown in Fig. 5.44. The waveforms explaining its operation are shown in Fig. 5.45.

Figure 5.46*a* shows another form of a pulse-counting detector. It consists of a comparator (a threshold circuit), a monostable or single-shot multivibrator, and an integrator. The comparator built around an operational amplifier, *Op-Amp*$_1$ (see Fig. 5.46*b*), converts the input signal (Fig. 5.47) to a train of voltage pulses, v_2, whose duration is limited by the zero-crossings of the input FM signal. The leading edges of these pulses trigger the monostable multivibrator built around NOT and AND logic gates coupled by a time-controlling R_1C_1 network. The output pulses of the monostable multivibrator (v_3 in Fig. 5.47) have a constant duration and height, and their repetition rate varies in direct proportion to the frequency variation of the input signal. The integrator built around a second operational amplifier, *Op-Amp*$_2$, averages the pulse train, thus generating a voltage, v_4, which varies in direct proportion to the pulse repetition rate.

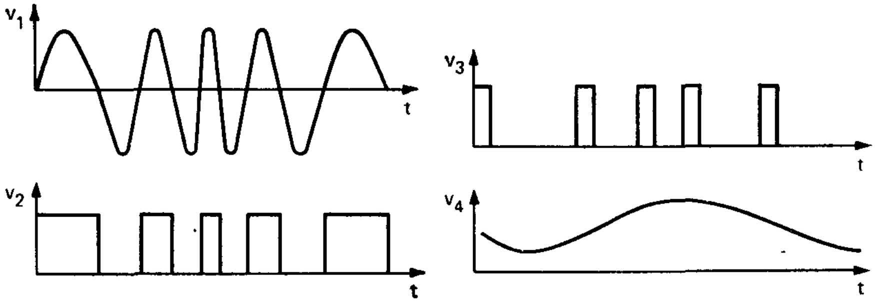


Fig. 5.47

Among the advantages offered by pulse-counting detectors are high quality of detection, independence of detection from variations in the centre frequency of the input signal, and availability in single IC packages.

Chapter Six

Manual and Automatic Receiver Controls and Indicators

6.1. Purpose and Types of Manual and Automatic Controls

A receiver will typically have several controls which vary in number and function according to its purpose and complexity. These include a control for tuning in a desired station, and controls to match the signal level and other signal parameters to the user's needs. Controls may be manual and automatic. Automatic control is effected in response to commands stored in a programmable control unit. With such an arrangement, nothing or very little is left for a human operator to do, except hitting a key to start the control or some other unit.

A receiver may operate under varying conditions. There may be variations in signal levels from different stations. Even the signal sent out by the same transmitter may vary in strength because of changes in the conditions of radio wave propagation. Furthermore, the signal frequency may be anything but stable because of transmitter instability or the Doppler effect. The signal frequency may undergo variations in the i.f. section of a receiver because the local oscillator used in the frequency converter does not supply a stable frequency. Reception conditions may also vary due to nonstationary additive and multiplicative noise. In all of such cases, one has to adjust the units, subassemblies and circuits of the receiver so as to obtain optimal reception.

Control may be effected locally or remotely. With remote control, the operator or the control unit is located at some distance from and linked to the receiver by telecontrol and teleindication facilities.

Manual control permits the use of electromechanical devices. For example, a receiver would until quite recently be tuned to the desired frequency mainly by selecting appropriate tuned-circuit inductors with a contact-type band selector switch and a continuous rotation of the tuning capacitor. With the advent of varactors instead of tuning capacitors, contact-type potentiometers came to be used to set the tuning voltage. Unfortunately, the use of electromechanical devices for remote or automatic control involves the use of specially designed electric motors, and this complicates receiver construction and impairs reliability. That is why remote and automatic control necessitates, as a rule, a change-over to purely electronic devices.

Automatic control is furthermore essential for reliable reception under rapidly varying conditions when a human operator would not be able to act with sufficient agility and accuracy if he were left with manual controls alone. No less important, automation simplifies the operator's functions or even makes human attendance completely unnecessary.

Control functions are especially exacting in a complex situation when a need arises for reliable reception of complex signals under varying transmission conditions and in the presence of various forms of noise. The adaptation of a receiver to such conditions in order to provide a most faithful reproduction of incoming information is a formidable task. A human operator tackles it by trial and error, which is a time-consuming procedure often entailing the loss of some information. This task is handled far better by automatic electronic controls which rely for their functioning on high-speed microprocessors.

A major trend in the development of all technology, including radio communications and broadcasting, is to build and use telecontrolled and fully automated systems. With them, all control functions essential for a given piece of equipment to meet the applicable performance specifications are carried out automatically.

The automatic controls most commonly used in receivers are *automatic gain control* (AGC) and *automatic frequency control* (AFC).

AGC holds the i.f. amplifier output substantially constant and sufficiently high for faithful reproduction of messages coming in from stations differing in transmitter power, located at different distances from the receiver, and operating under varying conditions of radio wave propagation. Being very simple, AGC is used in almost all receivers.

AFC has as its objective to hold the received signal spectrum located optimally within the receiver bandwidth in the face of the variations caused in the transmitter frequency and receiver-circuit tuning by various factors. AFC is used in almost all tyres of professional receivers and in many broadcast receivers.

Strong noise may impair reception or even render it completely unreliable. A need may arise to adjust receiver circuits not only in terms of frequency, gain, or signal level, but also for maximum reliability of receiver information. Among other things, this purpose can be served by automatic selectivity control effected by varying the bandwidth and the frequency response of the receiver. With strong signals or a low noise level, the bandwidth can be extended, thus improving message reproduction. With weak signals or a high interference level a reduced bandwidth may mitigate the detrimental effect of noise although the reception of wanted signals will be of a lower quality as compared with the previous case. Therefore, the objective of automatic control is to maintain an optimal bandwidth such that the receiver will reproduce the received information with a minimum of loss. Such automatic controls, along with some others, are used more seldom than AGC and AFC since the conditions in which they would yield satisfactory results are less definitive.

In our subsequent discussion we will be concerned solely with electronic systems because they are simpler in design, more reliable and faster than their electromechanical counterparts.

6.2. Automatic Gain Control: Types and Characteristics

When the voltage at the input to an amplifier is a minimum, $V_{\mathrm{in,min}}$, the amplifier gain must be a maximum, K_{max}, so that the output voltage, $V_{\mathrm{out,min}}$, is sufficient for normal reproduction of the incoming message. Here, $V_{\mathrm{in,min}}$ defines the receiver sensitivity. As the input voltage rises, the gain must automatically be brought down so as to maintain the output voltage substantially constant. Automatic gain control will do this job as required if the input and output voltages are related such that $K = V_{\mathrm{out}}/V_{\mathrm{in}}$. This relation is illustrated by curve *1* in the plot of Fig. 6.1.

Ordinarily, it is not required for the output voltage to be exactly constant. To simplify AGC, the output voltage is allowed to vary over a range such that there is no appreciable overloading of the receiver circuits and signal distortion. As the input voltage rises to $V_{\mathrm{in,max}}$, the output voltage goes up to $V_{\mathrm{out.max}}$, and the minimum gain in such a case is $K_{\rm min} = V_{\rm out,max}/V_{\rm in,max}$. This relation is illustrated by curve *2* in the plot of Fig. 6.1, where it runs slightly above curve *1*.

If the signal voltage at the input to the amplifier is below $V_{\rm in,min}$, no normal reception will be possible because receiver noise would mask the desired signals. Still, the shape of the AGC characteristic above point A which corresponds to $V_{\rm in,min}$ and $V_{\rm out,min}$ is of interest to the receiver designer.

In the simplest case, what happens in the AGC loop as $V_{\rm in}$ falls below point A is a uniform extension of curve *1*; this portion of the hyperbola is shown dashed. In this type of AGC, often called *simple* or *normal AGC*, the control law will remain unchanged even if the input signal falls below the sensitivity threshold. It is not used for the following reasons.

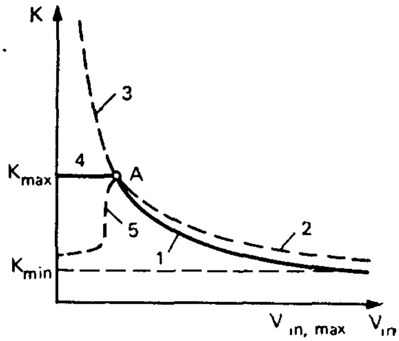


Fig. 6.1

1. When $V_{\rm in} < V_{\rm in,min}$, the voltage at the amplifier output will remain substantially constant, but it will be a mixture of the wanted signal and noise, with the noise contribution increasing as $V_{\rm in}$ decreases.

2. For the gain to be raised above $K_{\rm max}$ along curve *3*, one would have to include additional amplifying stages in the receiver. Yet, this increase in gain would be useless if not detrimental because there would be noise present at the receiver output even if there is no signal coming from its source.

A way out of the above situation is to hold AGC disabled so long as the input voltage remains below $V_{\rm in,min}$. Then the gain to the left of point A remains constant and equal to $K_{\rm max}$, as represented by curve *4*. The action of AGC is 'delayed' until the input voltage has reached $V_{\rm in,min}$; from that instant on, the AGC circuit maintains the output voltage as constant as required. Quite aptly, this type of control is called *delayed AGC*. Given the same quality of gain control above $V_{\rm in,min}$, a receiver with a delayed AGC circuit is simpler in construction than one with simple AGC.

While a receiver incorporating an AGC loop is being tuned from one station to another, when there is no wanted signal present at its input, the gain is a maximum, which is why there is a maximum amplification of receiver and external noise. In broadcast receivers the AGC loop is modified so that no noise can reach the receiver output in the course of tuning, thus giving what is known as *quiet tuning*. For this purpose, the gain to the left of point A is brought down as shown by curve *5* in Fig. 6.1. This is *quiet AGC*.

For its operation, an electronic AGC loop needs a control voltage which could then be applied to the controlled circuits and change the gain as shown in Fig. 6.1. Since the action of the AGC loop depends on the signal voltage, the simplest way to obtain the desired control voltage is to use the rectified voltage of the received signal. If the available voltage is not sufficient, an additional amplifier is included in the AGC loop. The control voltage can be supplied by an amplitude detector. The requirements for such a detector are, however, different from those which apply when it is used for AM reception and serves to reproduce the envelope of the signal wave, as shown in Fig. 5.1. If the voltage of such a detector were fed to the AGC loop, the rise in the signal amplitude in step with modulation would cause the gain to decrease and vice versa. As a result, the signal voltage at the amplifier output would remain practically constant in amplitude, which means that the AGC loop would suppress the modulation of the received signal. Of course, this is objectionable because it is the modulation that carries the desired information.

To avoid modulation suppression, the control voltage should be free from the alternating component associated with modulation. This goal can be achieved in any one of two ways.

1. The time constant of the C_LR_L network at the detector output (see Fig. 5.2*a*) is increased so that the voltage across C_L remains approximately equal to the maximum amplitude of the detected voltage. This process is similar to that shown in Fig. 5.11, but the capacitor discharges at a slower rate. The output voltage tracks the peaks of the signal amplitude and does not reproduce the envelope of the amplitude. This is what is known as the *peak detector.*

2. The R_LC_L time constant answers the condition for distortionless detection, as defined by Eq. (5.35), and the voltage thus derived can be used to reproduce the received message; to this end, its alternating component is extracted with the aid of a d.c. blocking capacitor (see Fig. 5.13). On the other hand, the direct component (the averaged voltage) is utilized in the AGC loop which contains a low-pass filter. The filter blocks the passage of the alternating component, and the control voltage is thus proportional to the mean amplitude of the signal.

In the case of delayed AGC, the detector operates only when the signal voltage exceeds some threshold value. Below the threshold, the control voltage ought not to change the gain, which requirement is easy to satisfy if there is simply no control voltage present. This property exists in a diode detectôr biased to cutoff by a direct bias voltage, as shown in Fig. 6.2*a*. It is seen from Fig. 6.2*b* that the detector will remain disabled so long as $V < E_b$, that is, when there is no control voltage, V_c, available for the AGC loop.

In agreement with the foregoing, an AGC loop may contain the following sections of a receiver:

(a) r.f. and i.f. amplifiers adapted for gain control by variations in the control voltage;

(b) detectors to derive control voltages by signal rectification;

(c) additional amplifiers to boost the control voltage in cases where it is necessary to enhance the AGC efficiency;

(d) circuits supplying the threshold voltage for delayed AGC;

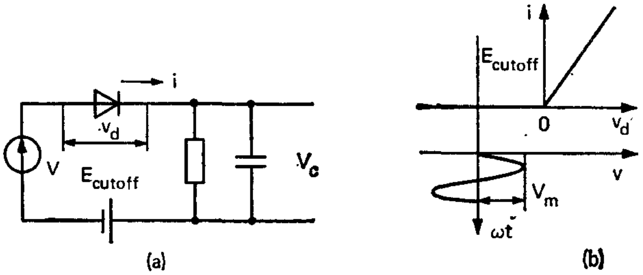


Fig. 6.2

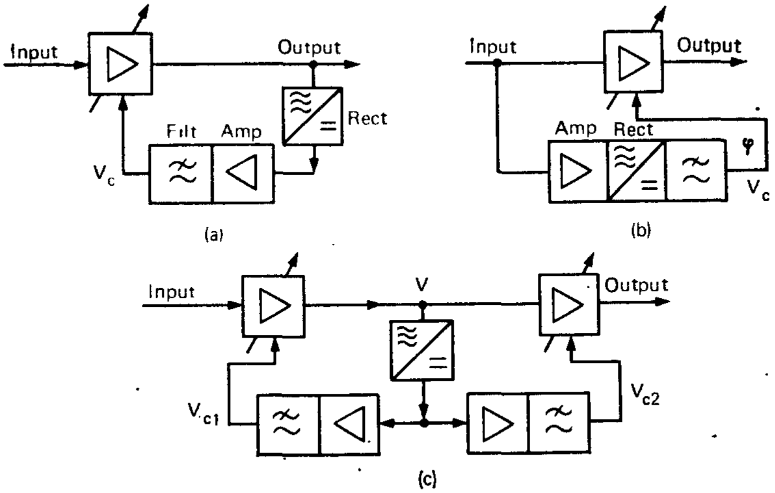


Fig. 6.3

(e) low-pass filters to suppress signal modulation products in the control-voltage circuits.

Three typical examples of AGC loops in simplified form and without delay circuits are shown in Fig. 6.3.

In the arrangement of Fig. 6.3*a*, the control voltage is derived by rectifying the amplified signal voltage appearing at the amplifier output. The voltage picked off the rectifier, *Rect*, is fed via an additional amplifier, *Amp*, and a low-pass filter, *Filt*, in a direction opposite to the signal flow in the controlled amplifier. At its output, it is fed to the preceding amplifier stages, for which reason this arrangement is known as *reverse AGC*. As an alternative, the amplifier, *Amp*, may be placed ahead of the rectifier. No amplifier may be used if the voltage at the output of the controlled amplifier is sufficiently high.

With reverse AGC, the gain is controlled owing to changes in the control voltage, V_C, which in turn varies due to changes in the signal voltage at the output of the controlled amplifier. Thus, in the case of reverse AGC the output voltage inevitably and of necessity varies. With a proper choice of the AGC loop parameters, however, such variations will not exceed an allowable limit.

In the AGC loop of Fig. 6.3*b*, the control voltage is derived as the input-signal voltage is amplified and rectified, and it acts in the same, 'forward' direction as the received signal flows in the controlled amplifier. This is what is known as *forward AGC*. In contrast to reverse AGC, the control voltage in a forward AGC loop is independent of the amplifier output voltage, which means that there is at least a theoretical possibility for the output voltage to be held fully constant. In practice, however, this possibility is never realized. As has been shown, the condition for the output voltage to be constant requires that the gain should vary in a precisely definite manner in response to variations in the input voltage (curve *1* in Fig. 6.1).

In practice, the gain is controlled by circuits whose properties vary with the control voltage. This relationship could be provided by nonlinear elements, but their characteristics strongly depend on the complex physical events occurring in them, and the curve shape can be controlled only to a very slight degree. If the function K (V_{in}) is represented by a drooping curve not unlike the theoretical curve in Fig. 6.1 and, with a suitable choice of circuit parameters, it merges with the latter at several points, the discrepancy between the two curves will inevitably be considerable within most of the portions between the points. In consequence, the output voltage will be anything but constant. The output voltage may deviate from the desired value by any amount, however large, but the AGC system will not respond to the fact, and the difference will remain.

Forward AGC becomes increasingly more difficult to implement if the voltage at the input to the controlled amplifier is likely to vary by a factor of several hundred or even thousand. For the control voltage to be able to act on the controlled amplifier, beginning with relatively weak input signals, the gain of the AGC amplifier (*Amp* in Fig. 6.3*b*) must be considerable, that is, of the same order of magnitude as the gain of the controlled amplifier. Unfortunately, with a substantial increase in the input voltage, the AGC amplifier will inevitably be overloaded and its nonlinearity will be accentuated.

In order to mitigate these events, the AGC amplifier itself needs an AGC loop of its own. Thus, a forward AGC loop is appreciably more complicated in construction than a reverse AGC loop and, still worse, it does not provide satisfactory gain control. For these reasons, forward AGC is not employed in the form described just above. Still, it may be useful as part of what is known as *mixed AGC* and shown in Fig. 6.3*c*. In this combination, the main job is done by reverse AGC. The controlled amplifier is divided into two sections, so that the major portion of amplification is provided by the first section, whereas the second section supplies only a low amplification. This section may, for example, be the final amplifying stage. The control voltage, V_{C1}, is derived by rectifying the voltage appearing at the output of the first section and is used to effect reverse gain control. The requirements for gain control quality here are rather lenient, which means that the first-section output voltage, V, is allowed to vary manyfold. This simplifies the implementation of gain control. At the same time, this voltage is utilized to produce a second control voltage, V_{C2}, which is used for forward gain control in the second section. Since this section is required to raise the gain only severalfold, the discrepancy between the theoretical and real gain control characteristics will not lead to marked variations in the output voltage of the second section. Also, the AGC loop uses the signal after it has been amplified by the first section, so there is no need for an additional high-gain amplifier such as is needed in the previous case.

While it is only slightly more elaborate in design than the arrangement in Fig. 6.3*a*, mixed AGC provides for a better gain control quality.

6.3. Methods of Gain Control

According to Eq. (3.95), the gain of an amplifying stage which contains a frequency-selective filter is given by

$$K \approx mn \mid Y_{21} \mid \rho K_{\mathrm{F}}$$

The filter gain, K_{F}, at a given signal frequency depends on the parameters of the resonant circuits that make up the filter, and the coefficient of coupling between these circuits. In consequence, the gain could be controlled in any one of several ways, as follows.

1. By varying the transconductance, Y_{21}, of the amplifying device. This is a feasible approach because the transconductance depends heavily on the direct voltages maintained at the electrodes of the amplifying device.

2. By varying the tapping-down factors, m and n. The problem would be simple to tackle, if connection were via a capacitive divider (see, for example, Figs. 2.17, 2.21, and 3.19). Then the tapping-down factor could be adjusted by varying the capacitances, which would be easy to do if varactors were used as variable capacitors. However, changes in capacitance would entail changes in the tuning of the resonance circuit and, as a consequence, an impairment in the frequency response and selectivity of the amplifier. That is why this approach is not used.

3. By varying the characteristic impedance, ρ. This would necessitate a change in the inductance and capacitance in mutually opposite directions, which is difficult to effect and would de-tune the resonant circuits. Therefore, this technique is not used, either.

4. By varying the filter gain, K_F. The required adjustment can be effected in any one of three ways, as follows:

(a) by varying the amount of damping in the tuned circuits through connection of variable-resistance networks. An increase in damping leads to a decrease in gain, but impairs selectivity. Since the gain has to be brought down when receiving strong signals, the decreased selectivity may be tolerated. This procedure is easy to carry out, and so it is used to some extent;

(b) by varying the amount of coupling between the tuned circuit of the bandpass filter. This technique is simple to implement in the case of capacitive coupling (see, for example, Fig. 3.28), with varactors used as coupling capacitors. However, the resultant range of changes in gain turns out rather small and, on top of that, changes are brought about in the frequency response and resonance frequency of the filter;

(c) by staggered tuning of the resonant circuits. This can readily be done with varactors used as the tuned-circuit capacitors. However, this approach entails changes in the frequency response and an impairment in selectivity because, with staggered tuning, the gain at the frequencies of likely noise signals may turn out greater than it is at the wanted-signal frequency.

As follows from the foregoing, preference should be given to the technique outlined in (1), namely, transconductance adjustment. Since it involves variations in the electrode voltages of the associated amplifying device, this approach might be called voltage gain control.

It is to be noted that this technique entails some changes in the resonant frequency and frequency response of the amplifying stage. This happens because a change in the electrode voltages of an amplifying device brings about changes in its input and output impedances and, in consequence, in the impedances that are coupled into the associated tuned circuits. The resistive component of the coupled-in impedance affects the damping factor of the tuned circuit, and the reactive component affects its resonant frequency.

The frequency-related properties of an amplifier can be stabilized by incorporating variable-gain networks which would not affect its frequency response and its tuning frequency. As a rule, such networks are variable-ratio electronic attenuators.

There are other types of gain control, such as the pulse type. One example is illustrated in Fig. 6.4. The signal (see Fig. 6.4*a*) is passed through a chopper (which is essentially a transistor or a diode that is caused to turn on and off at a predetermined rate). This converts the incoming signal into a train of pulses with their duty factor (that is the ratio of pulse duration to pulse spacing) determined by control pulses fed to the switch circuit.

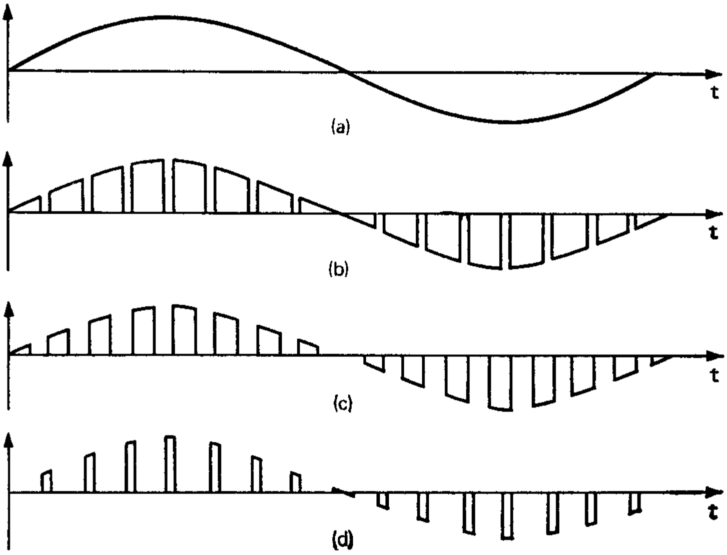


Fig. 6.4

The modulated signals thus obtained are then smoothened by a low-pass filter to obtain a voltage proportional to the average signal voltage or current which is a function of the pulse duty factor. With short pulses (see Fig. 6.4*d*), this voltage is many times lower than it is when the spacing between pulses is short (see Fig. 6.4*a*).

This type of control is highly efficient, but it should be used with care because short pulses have a broad frequency spectrum, and its high-frequency components and conversion products may interfere with signal reception, should they fall within the passbands of the r.f. and i.f. sections of a receiver.

An example of another gain control technique is one which utilizes the dependence of the conversion transconductance of the frequency converter (mixer) on the voltage supplied by the local oscillator (see Fig. 4.13). The conversion transconductance and, in consequence, the gain can be varied at will by varying the amount of coupling between the mixer and the local oscillator.

The variables that characterize the performance of a variable-gain circuit is the gain-control ratio, γ, defined as the ratio of the maximum gain, K_{max}, to the minimum gain, K_{min}. The required value of γ depends on the range of changes in the input signal voltage, V_{in}, and in the output signal voltage, V_{out}. If we denote

and
$$\left.\begin{array}{r} V_{\text{in, max}}/V_{\text{in, min}}=\alpha \\ V_{\text{out, max}}/V_{\text{out, min}}=\beta \end{array}\right\} \qquad (6.1)$$

then, noting that in a receiver with AGC.

$$K_{\text{max}} = V_{\text{out,min}}/V_{\text{in,min}}$$

and

$$K_{\text{min}} = V_{\text{out,max}}/V_{\text{in,max}}$$

we get

$$\gamma = (V_{\text{out,min}}/V_{\text{in,min}})/(V_{\text{out,max}}/V_{\text{in,max}})=\alpha/\beta \qquad (6.2)$$

The gain-control ratio, γ, must be fairly high. If, for example, a broadcast receiver should be capable of receiving signals with $V_{\text{in,min}} = 1$ µV and, when tuned to a nearby high-power station, it should be able to receive signals with $V_{\text{in,max}} = 20$ mV without overload-caused distortion, then

$$\alpha = (2 \times 10^{-2})/(1 \times 10^{-6}) = 2 \times 10^{4}$$

Suppose that the maximum-to-minimum output voltage ratio is $\beta = 2$. Then the gain-control ratio will be

$$\gamma = (2 \times 10^{4})/2 = 10^{4}$$

Ordinarily, a single amplifying stage cannot change the gain by a factor of more than a few tens. As a rule, it is difficult to reduce the gain below some certain limit since the r.f. signal almost inevitably 'creeps' through stray capacitances, mutual inductances, and conductances. Also, in order to reduce the gain of nonlinear elements by an appreciable amount, one would have to operate such elements under conditions in which their nonlinearity is felt especially strong and may cause signal distortion. The gain has to be brought down when the signal is strong enough for the nonlinearity to show up, in which case the nonlinear conversion products proportional to the second and higher powers of the signal amplitude would be especially troublesome. For this reason, gain control is effected in the stages located nearer to the amplifier input where the received signal has not yet been amplified very much. Unfortunately, gain control in the input stages of a receiver might impair selectivity. It is in the early stages that the interference from nearby stations operating at closely-spaced frequency channels is felt especially strong. In the presence of nonlinear elements, such an interference cannot be removed owing to cross modulation and intermodulation (see Sec. 1.8). In the succeeding stages, spurious responses are attenuated by the resonant circuits, and nonlinear processes are less detrimental.

If it is desired to obtain a high overall value for γ with a limited gain-control ratio of the individual stages, the AGC loop may be extended to include several stages. Since the gains of cascaded stages are multiplied together, their gain-control ratios with likewise be multiplied together, that is

$$\gamma = \gamma_1\gamma_2\gamma_3 \cdots \gamma_n$$

For higher efficiency, it is often practised to combine several different forms of gain control.

6.4. Examples of Practical AGC Circuits

Examples of AGC loops in which gain control is achieved by varying the transconductance of transistors are shown in Fig. 6.5. In the

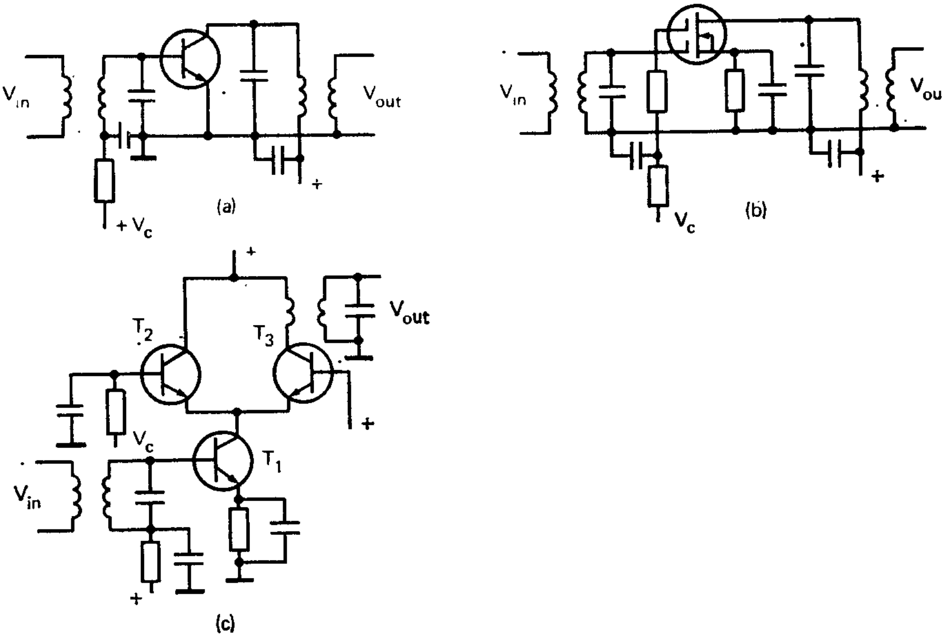


Fig. 6.5

circuit of Fig. 6.5*a*, the control voltage is fed to the transistor base. The transistor transconductance and, as a consequence, the stage gain are brought down as this voltage is lowered. In the circuit of Fig. 6.5*b*, the control voltage is fed to the second gate of a FET and likewise serves to change the FET transconductance.

In the circuit of Fig. 6.5*c*, the control voltage causes the current flowing through T_1 to be re-distributed between T_2 and T_3. As the current through T_2 increases, the current in T_3 decreases, and vice versa. As the current dividing into T_3 decreases, it causes a decrease in the alternating component produced by the signal voltage at the amplifier input, which is equivalent to a decrease in the transconductance.

Gain control by shunting the tuned circuit calls for the use of a resistive element controlled by variations in voltage or current.

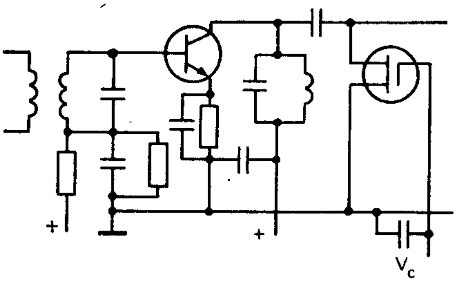


Fig. 6.6

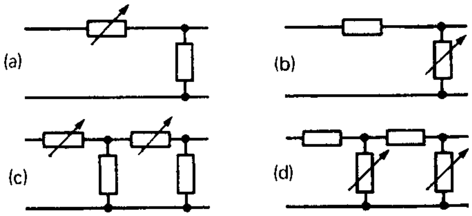


Fig. 6.7

Such an element may be a diode whose differential (dynamic) resistance is tens of ohms in the forward current region and hundreds of kilohms in the reverse current region. Within the portion where a change-over from forward to reverse current occurs, the diode characteristic is highly nonlinear, and this might bring about distortion in AM signals.

There are variable-resistance elements which remain sufficiently linear at alternating voltages as high as tens of millivolt. Apt examples are the collector-emitter circuit of a BJT and the drain-source circuit of a FET, as shown in Fig. 6.6.

An AGC loop may be based on a single-section attenuator, as shown in Fig. 6.7*a* and *b*, a two-section attenuator, as shown in Fig. 6.7*c* and *d*. If the variable-resistance element is a diode, the characteristic is most of all nonlinear within the transition portion from the low-ohmic forward current region to the high-ohmic reverse current region. Nonlinear distortion is especially pronounced if the diode is operated within this region when receiving strong signals. Therefore, preference should be given to attenuators in which the diode operates within the forward' current region in the case of strong signals. This requirement is best satisfied by attenuators with shunt controlled arms, such as shown in Fig. 6.7*b* and *d*. In the case of strong signals, the larger proportion of the input voltage is dropped across the resistors in the series arms and the voltage across the diodes is low, which fact serves to mitigate nonlinearity. In the case of weak signals, the control voltage serves to raise the diode resistance and gain.

The range of gain control can be extended by varying the resistance of both the shunt and series arms in the mutually opposite directions. An example of such an attenuator is shown in Fig. 6.8.

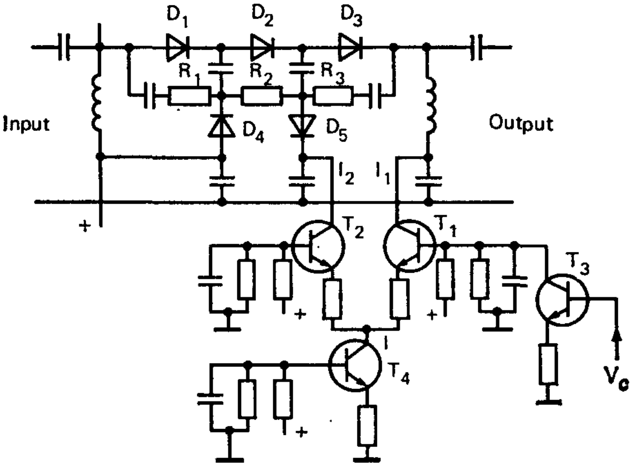


Fig. 6.8

Referring to the figure, the diodes D_1, D_2 and D_3 in the series arms of the attenuator are traversed by current I_1 of transistor T_1, whereas the diodes D_4 and D_5 in the shunt arms are traversed by current I_2 of transistor T_2. Current I_1 is controlled by voltage V_c via transistor T_3. The current $I = I_1 + I_2$ flowing through transistor T_4 is divided differently, depending on the value of V_c: when I_1 rises, I_2 falls, and vice versa. Therefore, when the resistances presented by D_1, D_2 and D_3 decrease, those of D_4 and D_5 increase, and the gain goes up. If, in contrast, the resistances presented by D_1, D_2 and D_3 increase, those of D_4 and D_5 decrease, and the gain falls.

Resistor R_2 is provided to pass current I_2; resistors R_1 and R_3 make the circuit symmetrical and balanced.

6.5. Basic Variables and Characteristics of Reverse AGC

The principal objective in synthesizing an AGC loop is to select a method for gain control and to determine how the gain will vary with the control current or, which is more often the case, with the control voltage, V_c. In approximate form, this relationship is illustrated by the plot of Fig. 6.9. The required range of gain variation, stated in terms of the ratio γ (see Sec. 6.3), may be very wide, so the gain is laid off as ordinates in the plot of Fig. 6.9 on a logarithmic scale.

Referring to the plot of Fig. 6.9, one can find the required maximum control voltage, $V_{c,max}$. To do this, one assumes the maximum gain and finds the gain at the maximum signal voltage at the receiver input

$$K_{min} = K_{max}/\gamma$$

where γ is found by Eq. (6.2). Once K_{min} has been determined, it is an easy matter to find $V_{c,max}$ from the plot of Fig. 6.9.

Suppose that this voltage is generated in a reverse AGC loop by a diode detector such as shown in Fig. 6.2. In contrast to Fig. 5.9

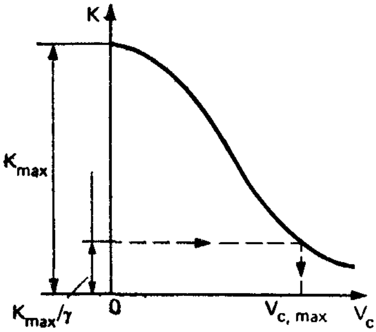


Fig. 6.9

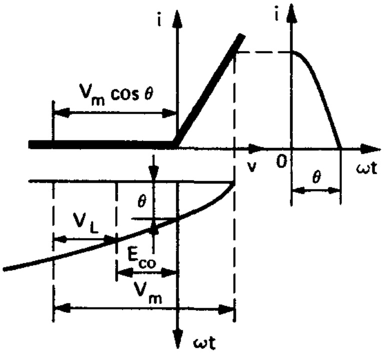


Fig. 6.10

which illustrates the operation of a diode into a load but without a cutoff bias voltage, E_b applied and to Fig. 6.2 where $V_m < E_b$ so that the diode is rendered non-conducting, the present case is illustrated by Fig. 6.10. Both V_L and E_b are applied to the diode, such that

$$V_L + E_b = V_m \cos\theta$$

Hence,

$$V_L = V_m \cos\theta - E_b$$

As V_m varies, the angle of current flow (the operating or conduction angle) changes somewhat, but this change is usually neglected because cos θ remains equal close to unity.

If the AGC loop is set up as shown in Fig. 6.3*a*, the output voltage of the amplifier, *Amp*, is fed to the detector directly. Let the amplifier gain be denoted as K_{amp}. Then

$$V_{c,max} = K_{amp}V_L = (V_{out,max} \cos\theta - E_b)\, K_{amp} \qquad (6.3)$$

where $V_{out,max}$ is the maximum amplifier output voltage applied to the AGC detector.

The cutoff bias voltage E_b must be equal to the minimum amplifier output voltage, $V_{out,min}$, so that until the latter is reached, the AGC loop will remain disabled (see Fig. 6.2*b*). Therefore, in view of Eq. (6.1),

$$V_{c,max} = V_{out,\ min} (\beta \cos \theta - 1) K_{amp}$$

Accordingly, the required amplifier output voltage is

$$V_{out,min} \geqslant V_{c,max}/K_{amp} (\beta \cos \theta - 1)$$

If the above condition is not satisfied at $K_{amp} = 1$, an additional amplifier is needed, with a gain

$$K_{amp} \geqslant V_{c,max}/V_{out,min} (\beta \cos \theta - 1)$$

In contrast to operation of the amplitude detector examined in Sec. 5.5, the case at hand does not require that there should be no nonlinear distortion. Therefore, the detector load resistor, R_L, may have a sufficiently high value so that $\cos \theta \approx 1$.

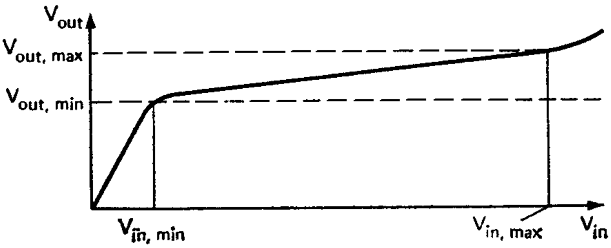


Fig. 6.11

An idea about the efficiency of AGC can be gleaned from a plot relating the controlled amplifier output voltage V_{out} to the receiver input voltage V_{in}. As follows from Eq. (6.3) applied to intermediate values,

$$V_c = \varphi (V_{out})$$

A plot of $K = \psi (V_c)$ is constructed in Fig. 6.9. Therefore,

$$V_{out} = K V_{in} = V_{in} \psi [\varphi (V_{out})]$$

The above equation can be solved for V_{out} only graphically because the function $\psi (V_c)$ defined in graphical form has no analytic solution. It is more convenient to proceed from the expression for the inverse function:

$$V_{in} = V_{out}/\psi [\varphi (V_{ont})]$$

Assuming several values of V_{out}, we obtain

$$V_c = (V_{out} \cos \theta - E_b) K_{amp}$$

(at $V_{out} < E_b$, $V_c = 0$). Now, referring to Fig. 6.9, find the value of K corresponding to the found value of V_c and calculate the input voltage as

$$V_{in} = V_{ont}/K$$

The plot based on the data for $V_{out} > V_{out,min}$ is shown in Fig. 6.11. So long as $V_{out} < V_{out,min}$, the AGC loop remains delayed, so $K = K_{max} =$ const, which means that the initial portion of the characteristic answers the formula $V_{out} = K_{max}V_{in}$. For ease of reading the values of V_{in} which can vary between broad limits, it is laid off as abscissa on a logarithmic scale.

6.6. Transients in a Receiver with AGC

In developing an AGC circuit, it is essential to choose the configuration and element values for the low-pass filter (*Filt* in Fig. 6.3) in the controlled voltage circuit. Should this filter have an excessively long time constant, the AGC loop would be too sluggish to respond in time to fast rises or falls in the signal voltage. On the other hand, if the filter time constant were too short, the signal would be likely to be distorted.

Suppose that the input to a receiver is an AM signal of the form

$$V_{in} = V_{in,0}\,(1 + m \cos \Omega t)$$

This signal is amplified, detected by the amplitude detector of the AGC loop, and passes through the AGC filter where the modulation-frequency component is partly suppressed and may change phase. As a result, V_c will, to a first approximation, be varying as

$$V_c = V_{c,0}\,[1 + \mu \cos (\Omega t + \varphi)]$$

To simplify the analysis, we have neglected here the fact that the generation of the control voltage may be accompanied by nonlinear distortion in the modulating function. This, means that, apart from the component at the angular frequency Ω the spectrum of this voltage may contain components at frequencies 2Ω, 3Ω, etc.

When this voltage is fed to the controlled amplifier whose operation is characterized by a function of the form $K = \psi\,(V_c)$, shown in Fig. 6.9, the gain will, likewise to a first approximation, be varying as

$$K \approx K_0\,[1 - S_{amp}\mu \cos (\Omega t + \varphi)]$$

where K_0 = amplifier gain corresponding to $V_{c,0}$
$S_{amp} = -dK/dV_c$ = slope of the utilized portion of the characteristic plotted in Fig. 6.9

In our case, this portion is deemed to be linear, which means that in this case, too, we neglect nonlinear distortion.

On denoting $S_{\text{amp}}\mu$ as ν and on multiplying V_{in} by K, we find a more accurate expression for variations in the amplifier output voltage

$$V_{\text{out}} = V_{\text{in},0}K_0 (1 + m \cos \Omega t) [1 - \nu \cos (\Omega t + \varphi)]$$

or, in a different form,

$$V_{\text{out}} = V_{\text{in},0}K_0 [1 + m \cos \Omega t - \nu \cos (\Omega t + \varphi) - m\nu \times 0.5 \cos (2\Omega t + \varphi) - m\nu \times 0.5 \cos \varphi]$$

The above result suggests the following conclusions.

(a) Variations in the gain at the modulation frequency causes a change in the modulation factor of the signal. When $\varphi = 0$, the resultant modulation factor is $| m - \nu |$, which means that the modulation may be attenuated or suppressed.

(b) In addition to modulation by the fundamental angular frequency Ω, the signal is modulated by the second harmonic at angular frequency 2Ω, which means that modulation distortion takes place. The harmonic distortion factor is proportional to ν. The modulation will not be attenuated at $\nu = 0$, that is, if the alternating component of the detected voltage has been fully suppressed by the filter.

An improper choice of the filter configuration and element values might lead to gain instability, so that instead of being stabilized, the output signal voltage may suffer appreciable variations. In order to be able to ensure stability, it is important to know what nonstationary (transient) processes occur in the AGC amplifier. Consider the AGC circuit of Fig. 6.3*a*.

Ordinarily, the time constant for the detector load is chosen to be relatively short so that the load voltage settles at its steady value much faster than the voltage at the filter output. Such a choice of the time constant is warranted for diode detection because (see Fig. 5.14) the voltage across the detector load rises and falls at different rates. The capacitor charges via the diode which has a low internal resistance, and reaches its final charge rapidly as the amplitude of the applied alternating voltage increases.

When the amplitude decreases, the diode ceases conducting, and the capacitor discharges through the high-ohmic load resistor, which means that the discharge goes on at a markedly lower rate. In consequence, if transients in the detector were allowed to play an important role, the AGC circuit would operate differently during positive and negative changes in the signal voltage. To avoid this, it is usual to arrange for the detector to have a short time constant, in which case its transients may be ignored.

The output voltage of the smoothing filter activates the AGC loop. Its parameters change practically without a time lag, but this change entails a change in the voltage and current of the signal being amplified, thus leading to nonstationary events occurring in the tuned circuits and filters. However, it is legitimate to neglect these events, that is, to assume that the transients are solely associated with the AGC filter. This assumption is based on the fact that the r.f. and i.f. amplifiers along with the detector must reproduce the modulated wave and the modulating message, whereas the smoothing filter, as has been found, must suppress intermodulation products, which implies that it should be more sluggish in its response.

In a steady state,

$$V_{\text{out}} = KV_{\text{in}}$$

and the gain K is a function of the control voltage V_c:

$$K = \varphi\,(V_c)$$

In approximate terms (neglecting the cutoff bias voltage), the control voltage may be defined as

$$V_c = V_{\text{out}}K_d K_{\text{amp}} K_f$$

where K_{amp} = gain of the amplifier, *Amp* (see Fig. 6.3*a*)
K_d = voltage gain of the detector
K_f = gain of the filter

In a steady state, K_f is close to unity. To a first approximation, K_d may likewise be deemed constant and close to unity.

Now consider the relationships existing in the same system after the signal voltage has risen by a small increment, ΔV_{in}. Then the output voltage, V_{out}, will likewise rise by a small amount, ΔV_{out}, which fact causes the control voltage to increase by

$$\Delta V_c = \Delta V_{\text{out}} K_{\text{amp}} K_d K_f$$

The gain, too, takes on a new value

$$K' = \psi\,(V_c + \Delta V_c)$$

If ΔV_c is small, it is legitimate, on expanding K' into a series of powers of ΔV_c, to retain only the first terms of the series

$$K' \approx \psi\,(V_c) + [d\psi\,(V_c)/dV_c]\,\Delta V_c$$

On denoting, as before, $-dK/dV_c = S_{\text{amp}}$, let us write this expression as

$$K' = K - S_{\text{amp}}\,\Delta V_c$$

where S_{amp} is a positive factor characterizing the sensitivity of the amplifier towards variations in the control voltage, V_c. This factor can be found graphically from the plot of Fig. 6.9. Thus, we get

$$V_{\text{out}} + \Delta V_{\text{out}} = (V_{\text{in}} + \Delta V_{\text{in}})\,(K - \Delta V_{\text{out}} S_{\text{amp}} K_{\text{amp}} K_d K_f)$$

On subtracting $V_{\text{out}} = V_{\text{in}}K$, we obtain

$$\Delta V_{\text{out}} = \Delta V_{\text{in}} K - (V_{\text{in}} + \Delta V_{\text{in}})\,\Delta V_{\text{out}} S_{\text{amp}} K_{\text{amp}} K_d K_f$$

Hence,

$$\Delta V_{\text{out}} = V_{\text{in}}K/[1 + (V_{\text{in}} + \Delta V_{\text{in}})\, S_{\text{amp}}K_{\text{amp}}K_{\text{d}}K_{\text{f}}]$$

On setting $\Delta V_{\text{in}} \ll V_{\text{in}}$, the above expression may be simplified as

$$\Delta V_{\text{out}} \approx \Delta V_{\text{in}}K/(1 + S_{\text{amp}}V_{\text{in}}K_{\text{amp}}K_{\text{d}}K_{\text{f}}) \qquad (6.4)$$

The product $S_{\text{amp}}V_{\text{in}}$ depends on the efficiency of AGC. Let us write S_{amp} as the ratio of finite increments:

$$S_{\text{amp}} = \Delta K/\Delta V_{\text{c}}$$

Here,

$$\Delta V_{\text{c}} = \Delta V_{\text{in}}KK_{\text{amp}}K_{\text{d}}K_{\text{f,0}}$$

where $K_{\text{f,0}}$ is the filter gain in a steady state. Therefore,

$$S_{\text{amp}}V_{\text{in}} = \frac{\Delta K}{\Delta V_{\text{in}}KK_{\text{amp}}K_{\text{d}}K_{\text{f}}}\, V_{\text{in}} = [(\Delta K/K)/(\Delta V_{\text{in}}/V_{\text{in}})]\, (1/K_{\text{d}}K_{\text{amp}}K_{\text{f,0}})$$

In the case of high gain control efficiency, the relative change in gain is roughly proportional to the relative change in the input voltage. Therefore, the output voltage in a steady state remains nearly unchanged. If so, the ratio $(\Delta K/K)/(\Delta V_{\text{in}}/V_{\text{in}})$ is close to unity. Let us denote this ratio as K_{c} and let us call this quantity the *gain control quality factor*. It does not remain constant over the range of variations in the input voltage, but it may be deemed to have a definite value at a fixed initial input voltage. Therefore,

$$S_{\text{amp}}V_{\text{in}}K_{\text{d}}K_{\text{amp}}K_{\text{f}} = K_{\text{c}}K_{\text{f}}/K_{\text{f,0}}$$

The product $\Delta V_{\text{in}}K$ in the numerator of Eq. (6.4) is the increment in the output voltage which would have taken place if the AGC circuit had not responded to increments in the signal. Let us denote $\Delta V_{\text{out}}K$ as $\Delta V_{\text{out,0}}$. Then Eq. (6.4) will take the form

$$\Delta V_{\text{out}} = \Delta V_{\text{out,0}}/[1 + K_{\text{c}}\,(K_{\text{f}}/K_{\text{f,0}})] \qquad (6.5)$$

Let us denote $\Delta V_{\text{out}}\,(t)/\Delta V_{\text{out,0}}$ as $\xi\,(t)$ and write Eq. (6.5) as a Laplace transform equation* for transients

$$\xi\,(t) \rightarrow 1\,(p)/[1 + K_{\text{c}}K_{\text{f}}\,(p)/K_{\text{f,0}}]$$

Figure 6.12 shows the circuits of a single half-section, a double half-section, and a triple half-section filter. Assume that the filters are operating under practically in open-circuit condition and take it that in a steady state $K_{\text{f,0}}$ is approximately equal to unity in all the three cases. In a single half-section filter

$$v_2 = v_1/(1 + j\omega C_1R_1)$$

* Note that, as adopted in Soviet usage, the Laplace operator is symbolized as $p = \sigma + j\omega$, and ought not to be confused with the Heaviside operator $p = d/dt$.— *Translator's note.*

or, writing in Laplace-transform notation,

$$K_f(p) = 1/(1 + pa_1)$$

where $a_1 = R_1C_1$. In a double half-section filter,

$$K_f(p) = 1/(1 + pa_1 + p^2a_2)$$

where

$$a_1 = C_1R_1 + C_2R_2 + C_2R_1$$
$$a_2 = C_1R_1C_2R_2$$

In a triple half-section filter,

$$K_f(p) = 1/(1 + pa_1 + p^2a_2 + p^3a_3)$$

where

$$\begin{aligned} a_1 &= C_1R_1 + C_2R_2 + C_3R_3 + C_2R_1 + C_3R_1 + C_3R_2 \\ a_2 &= C_1R_1C_2R_2 + C_1R_1C_3R_2 + C_3R_3C_2R_2 \\ &\quad + C_3R_3C_2R_1 + C_1R_1C_3R_3 \\ a_3 &= C_1R_1C_2R_2C_3R_3 \end{aligned}$$

In the general case of a filter having n half-sections,

$$K_f(p) = 1\Big/\left(1 + \sum_{k=1}^{n} a_k p^k\right)$$

Therefore,

$$\xi(t) \to 1(p)\Big/\left[1 + K_c\Big/\left(1 + \sum_{k=1}^{n} a_k p^k\right)\right]$$

That, is,

$$(1 + K_c)\,\xi(t) \to 1(p)\left(1 + \sum_{k=1}^{n} a_k p^k\right)\Big/\left[1 + \sum_{k=1}^{n} p_k a_k/(1 + K_c)\right]$$

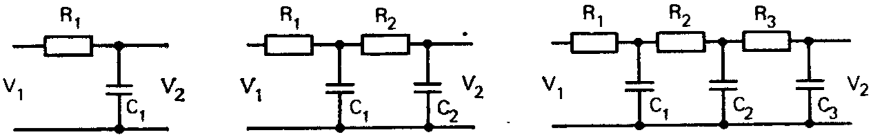


Fig. 6.12

This form of differential equation has an exponential solution, the coefficients of t in the exponents being the roots of the denominator, that is, the roots of the characteristic equation

$$1 + \sum_{k=1}^{n} p_k a_k/(1 + K_c) = 0 \qquad (6.6)$$

For a single-section filter, the characteristic equation has the form

$$1 + pa_1/(1 + K_c) = 0$$

Hence,

$$p = -(1 + K_c)/a_1 = -(1 + K_c)/C_1R_1$$

It follows then that the transient response will be an exponential and aperiodic one with a time constant

$$\tau = C_1R_1/(1 + K_c)$$

Referring to a table of Laplace transform pairs, we find that the Laplace transform equation in our case has the following solution:

$$(1 + K_c)\,\frac{\Delta V_{\text{out}}(t)}{\Delta V_{\text{out},0}} = 1 + K_c \exp\left[(-t/C_1R_1)(1 + K_c)\right]$$

A plot of the transient response for the case at hand is shown in Fig. 6.13 by the full curve. Just as the voltage changes stepwise, the AGC loop cannot bring down the gain at once, for which reason the initial increase in output voltage is the same as it was with the gain existing prior to the stepwise change. As the output voltage of the low-pass filter rises, the gain goes down, and the increment in the output voltage approaches its steady-state value, $\Delta V_{\text{out},0}/(1 + K_c)$.

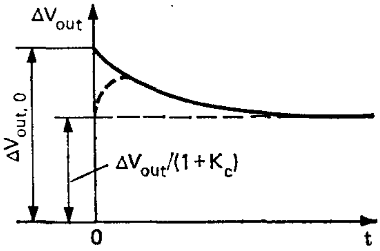


Fig. 6.13

What we have examined is in effect a hypothetical case. In practice, the input voltage of a receiver changes gradually rather than stepwise. Even in the case of a stepwise change, the output voltage of an amplifier would vary gradually because of the limited amplifier bandwidth. Therefore, the output voltage is incremented by ΔV_{out} as shown by the dashed curve in Fig. 6.13, that is, the transient increase ΔV_{out} over and above the steady-state value is smaller in magnitude. If follows from the foregoing that the time constant determining the speed of response of the AGC loop is by a factor of $1 + K_c$ shorter than that of the filter.

If the input voltage of the filter is changed stepwise by an amount equal to a steady-state value, the rate of change of the voltage at the output of the filter should, it would seem, change at a rate determined by its time constant, R_1C_1. Actually, however, the situation is different. At first, the gain has not yet been changed by AGC, and the increment in the voltage exceeds its steady-state value. For this reason, the voltage at the output of the filter will likewise rise and will, during the same span of time, approach the steady-state value faster, which is equivalent to a decrease in the time constant.

For an AGC loop containing a double-section filter, the characteristic equation, according to Eq. (6.6), has the form

$$a_2p^2 + a_1p + 1 + K_c = 0$$

Its roots are

$$p = -(a_1/2a_2) \pm [(a_1/2a_2)^2 - (1 + K_c)/a_2]^{1/2}$$

If $(1 + K_c)/a_2$ is greater than $(a_1/2a_2)^2$, these roots are complex, which corresponds to an oscillatory transient response. However, the real parts of the roots are negative, therefore, even if such an oscillatory response does occur, it dies out. Variations in the control voltage and, in consequence, in the gain have a detrimental effect on the quality of signal reproduction, therefore they are objectionable. The condition for the transient response to be aperiodic has the form

$$(a_1/2a_2)^2 > (1 + K_c)/a_2$$

or, on substituting for a_1 and a_2,

$$(\tau_2/\tau_1)^{1/2} + (\tau_1/\tau_2)^{1/2} [1 + (C_2/C_1)] > 2 (1 + K_c)^{1/2}$$

or, in a different form,

$$(\tau_1/\tau_2)^{1/2} + (\tau_2/\tau_1)^{1/2} [1 + (R_1/R_2)] > 2 (1 + K_c)^{1/2}$$

where $\tau_1 = C_1R_1$ and $\tau_2 = C_2R_2$ are the time constants of the filter sections.

If K_c is approximately equal to unity, then

$$2 (1 + K_c)^{1/2} \approx 2.8$$

which means that at, say, $\tau_1 = \tau_2$, the condition will not be satisfied for (C_2/C_1) less than 0.8 or for (R_1/R_2) less than 0.8.

A triple-section filter has a third-order characteristic equation. Its analysis shows that the circuit tends to be increasingly oscillatory in its response, so that the control voltage and, in consequence, the gain may experience undamped oscillations. This explains why it is not warranted to use more than two filter sections in a receiver.

6.7. Frequency Control and Phase-Locked Loops

Changes in the ambient conditions, notably temperature, entail changes in the element values of resonant circuits, especially in the capacitance of tuned-circuit capacitors and in the inductance of tuned-circuit coils. Still greater changes can occur secondary to changes in the capacitances of the electron devices connected to the tuned circuits. All of this causes a change in the tuning frequency and a shift in the frequency response. The r.f. signal spectrum and the initial position of the frequency response (shown by the full curve) can be seen in Fig. 6.14 which also shows (by the dashed curve) the position of the frequency response for an 'off-tune' condition. It is seen that the detuning or frequency offset may cause signal distortion or impede reception altogether. Also, it impairs selectivity because the signal is attenuated, and the adjacent-channel interference may fall within the passband of the i.f. amplifier.

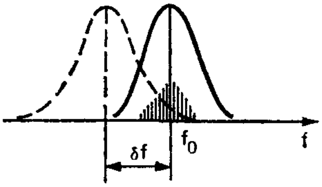


Fig. 6.14

With a satisfactory construction of the resonant-circuit components, the fractional detuning or the fractional frequency offset, $\Delta = \delta f/f_0$, may be of the order of one part in a thousand. On the other hand, the fractional bandwidth is of the order of 10^{-2}, which is by an order of magnitude greater than the likely frequency offset. In the circumstances, the frequency offset cannot be as large as shown in Fig. 6.14 and cannot entail the consequences we have mentioned. If we use highly stable (say, quartz-crystal) filters in the i.f. section, the matter of detuning may be disregarded completely.

Destabilizing factors affect the tuned circuit of the local oscillator as well and may cause its frequency to change by an amount of the same order

$$\Delta f_{LO} = \Delta_{LO} f_{LO}$$

If the i.f. of a receiver is $f_i = f_{LO} - f_s$, a change in the local-oscillator frequency will cause the i.f. to change by δf_i equal to δf_{LO}. The result of local-oscillator detuning will be the same as is shown in Fig. 6.14: the signal spectrum will be shifted away from the resonant frequency of the i.f. section. The fractional frequency offset or detuning will then be equal to

$$\Delta_i = \delta f_i/f_i = \delta f_{LO}/f_{LO}$$

Let us write it as

$$\Delta_i = (\delta f_{LO}/f_{LO})\,(f_{LO}/f_i)$$

Since $f_{LO} = f_s + f_i$, we then have

$$\Delta_i = \Delta_{LO}\,(1 + f_s/f_i)$$

Since f_i is a small fraction of f_s, the instability of the local oscillator leads to the frequency instability of the receiver which is much greater than the instability of the resonant circuits of the selective amplifier section.

The local oscillator can be stabilized by using a separate quartz resonator, or crystal unit, for each stabilized frequency, which obviously complicates receiver construction. State-of-the-art receivers use frequency synthesizers with a single reference crystal-controlled oscillator. This is known as *coherent frequency synthesis*. With this arrangement, a spectrum of highly stable frequencies, spaced 100 Hz apart (which is usually sufficient for practical purposes) can be produced.

In a continuously tunable receiver, the local oscillator is stabilized by an automatic control loop. When a receiver has several frequency converters, the control loop will "discipline" the local oscillator whose stability is most important. In some cases, additional control loops may be used for the remaining local oscillators.

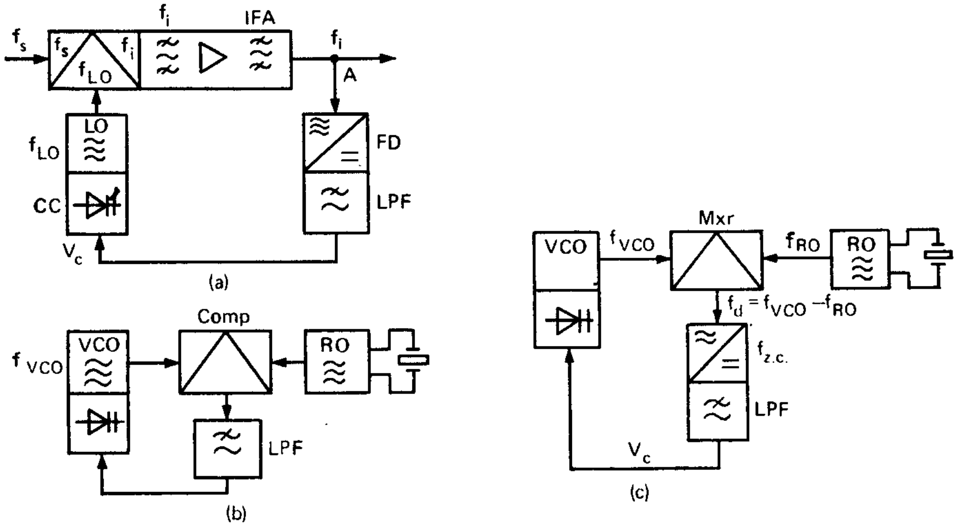


Fig. 6.15

The devices intended to effect frequency control may be classed in more than one way, as follows:

(a) according to the type of device used to supply the reference frequency with which the frequency of the controlled oscillator is comparèd;

(b) according to the parameters of the reference and controlled waves lying at the basis of automatic control.

In the systems falling in case (a), the frequency in a receiver may be compared with:

— a frequency at which the electronic circuit involved acquires some special properties: this may be a resonant frequency, the frequency at which a bridge circuit will be at balance, and so on;

— the frequency of a stable oscillator;

— reference frequencies of both tŷpes (mixed systems).

Examples of the respective types are shown in Fig. 6.15. In the arrangement of Fig. 6.15*a*, the reference frequency is the resonant frequency of a circuit which is part of a frequency detector. *FD*.

The frequency at which the response curve of the frequency detector crosses zero (see Figs. 5.33 and 5.39) is the frequency to which the i.f. amplifier is tuned. Whenever the local-oscillator frequency, f_{LO}, or the signal frequency, f_s, deviate from the value corresponding to the exact tuning, a change occurs in f_i. The output voltage of the frequency detector corresponds then to the direction and amount of change in frequency.

The output voltage of the frequency detector is passed through a low-pass filter, *LPF*. As is the case with an AGC loop, the purpose of the low-pass filter is to suppress the voltage variations that are caused as the carrier is modulated by the intelligence being transmitted. The control voltage, V_c, thus derived is then applied to the control circuit, *CC*, of the local oscillator so that f_{LO} is caused to vary in a direction to reduce the frequency offset.

In the arrangement of Fig. 6.15*b*, the frequency of a voltage-controlled oscillator, *VCO*, is compared with that of a reference oscillator, *RO*, in a circuit called the comparator, *Comp*. Whenever there is a difference between f_{VCO} and f_{RO}, a voltage appears at the output of the comparator. This voltage is passed through a low-pass filter, *LPF*, and goes to the control circuit, *CC*, which adjusts the voltage-controlled oscillator so as to minimize the difference in frequency.

In the mixed arrangement of Fig. 6.15*c*, the VCO and RO voltages with frequencies f_{VCO} and f_{RO}, respectively, are applied to a mixer, *Mxr*, which produces an output voltage with a difference frequency, f_d. This voltage is applied to a frequency discriminator, *FD*, with a "zero-crossing" frequency, $f_{z.c}$. When f_d deviates from $f_{z.c.}$, the frequency discriminator produces an output voltage which, on passing through a low-pass filter, *LPF*, is applied to the control circuit which adjusts the VCO so as to minimize the frequency offset.

Systems falling in case (b), that is, those differing in the parameter being compared, may be classed into those with frequency comparison and those with phase comparison. In the former case, the sensing element of the control loop is a frequency detector, as shown in Fig. 6.15*a* and *c*. This type of device is called the *automatic frequency control loop*, or AFC (afc) for short. In the latter case, that is, in a system with phase comparison, the comparison is carried out between the phases of what is called a voltage-controlled oscillator, *VCO*, and a reference oscillator, *RO*. In this case, the comparator is a phase detector. The output of the phase detector is a function of the phase difference between the two signals. This error voltage, after low-pass filtering in the loop filter, is applied to the modulation input of the VCO. in such a way that the VCO signal phase follows the reference signal phase. It is then said to be locked to the latter in phase. Hence the name '*phase-locked loop*', or PLL for short. This principle is implemented in the arrangement of Fig. 6.15*b*.

A PLL utilizes the fact that when the two applied signals differ in frequency (say, f and $f - \delta f$) and, as a consequence, in period (T and $T + \Delta T$), a varying phase shift arises between them. To demonstrate, the voltages

$$v = V \cos (\omega + \delta \omega) t$$

may be written as

$$v = V \cos (\omega t + \varphi)$$

where $\varphi = \delta \omega t$ is the varying phase shift. If, for example, the difference in frequency is 1 Hz, then in one second one signal will be shifted relative to the other by a cycle, or a whole period, which means that in one second the phase shift will change by one-tenth of a cycle, or by 36°. If we apply the two voltages to a phase detector, its output voltage may be very high for any arbitrarily small difference in frequency, alhough the phase is then varying at a low rate. This explains why a PLL can respond even to the minutest difference in frequency.

6.8. Automatic Frequency Control

Most often, an electronic automatic frequency control (AFC or afc) loop depends for its operation on a varactor connected in its tuned circuit, as shown in Fig. 6.15. The control voltage applied to the varactor brings about a change in the varactor capacitance and, as a consequence, in the frequency of the controlled oscillator. In some cases, there may be no need in oscillator frequency control at all. For example, the frequency of a transistor local oscillator may be controlled between narrow limits by feeding the control voltage to the supply circuit of the transistor. Since the local-oscillator frequency depends on the transistor capacitances which are lumped with the overall capacitance of the transistor tuned circuit, it will change because these capacitances are functions of the applied voltages.

Consider the AFC loop shown in Fig. 6.15*a*. Let the nominal value of the desired signal (or sending-station) frequency be $f_{s,0}$, and designate the local-oscillator frequency in the case of exact tuning as $f_{LO,\,0}$, and the resultant intermediate frequency, likewise at exact tuning, as $f_{i,0}$. Suppose also that for some reason the received station frequency has changed by an amount equal to Δf_s so that it is now

$$f_s = f_{s,\sigma} - \Delta f_s$$

where Δf_s may be positive or negative.

Assume further that as a result of instability and due to the action of the control voltage supplied by the AFC loop, the local-oscillator frequency has changed by an amount equal to Δf_{LO} so that it is now

$$f_{LO} = f_{LO,\,0} + \Delta f_{LO}$$

If f_{LO} lies above f_s, then the nominal values of the two frequencies are related as

$$f_{i,0} = f_{LO,0} - f_{s,0}$$

(If f_{LO} lies below f_s, then $f_{i,0} = f_{s,0} - f_{LO,0}$, but this does not affect the reasoning.) The intermediate frequency will likewise change to become equal to

$$f_i = f_{LO} - f_s$$

Let us denote the change in the intermediate frequency as Δf_i such that $\Delta f_i = f_i - f_{i,0}$. In the case at hand,

$$\Delta f_i = (f_{LO} - \cdot f_s) - (f_{LO,0} - f_{s,0})$$

Substituting here for f_{LO} and f_s, we get

$$\Delta f_i = \Delta f_{LO} + \Delta f_s$$

The deviation of the intermediate frequency from its nominal value by Δf_i causes a voltage to appear at the output of the frequency detector, thus activating the AGC loop which acts so that the local-oscillator frequency is changed in the reverse direction by an amount equal to $\delta' f_{LO}$. Therefore, if the initial change in the local-oscillator frequency is δf_{LO}, the AFC loop causes this frequency to become equal to

$$f_{LO} = f_{LO,0} + \delta f_{LO} - \delta' f_{LO}$$

The above expression is written on the assumption that the frequency detector voltage is of such a polarity that the frequency offset is minimized. In the opposite polarity, the local-oscillator frequency would change so as to increase rather than decrease the frequency offset. It follows then that

$$\Delta f_{LO} = f_{LO} - f_{LO,0} = \delta f_{LO} - \delta' f_{LO}$$

Accordingly,

$$\Delta f_i = \delta f_{LO} - \delta' f_{LO} + \Delta f_s$$

Let us denote $\delta f_{LO} + \Delta f_s$ as Δf. This is the overall frequency offset, or the total detuning, whose effect on the intermediate frequency should be removed by the AFC loop. In the general case, the constituent terms of Δf may take both like and unlike signs. In view of the notation we have adopted,

$$\Delta f_i = \Delta f - \delta' f_{LO}$$

The value of the correction term $\delta' f_{LO}$ is a function of the control voltage, V_c, supplied by the frequency detector, that is,

$$\delta' f_{LO} = \zeta\,(V_c)$$

Therefore,

$$\Delta f_{\mathrm{i}} = \Delta f - \zeta\,[\psi\,(\Delta f_{\mathrm{i}})]$$

If we could solve the above equation for Δf_{i}, the solution $\Delta f_{\mathrm{i}} = \eta\,(\Delta f)$ could be used to plot the AFC characteristic. Unfortunately, the function $\psi\,(\Delta f_{\mathrm{i}})$ is rather complex, whereas the function

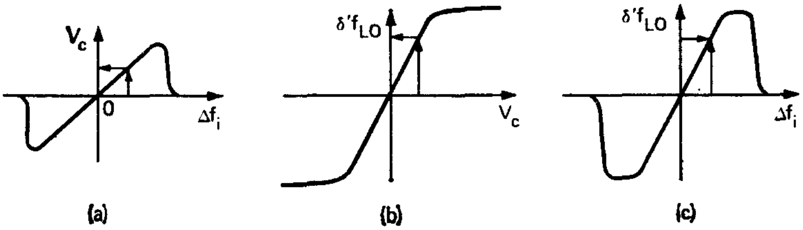


Fig. 6.16

$\zeta\,(V_{\mathrm{c}})$ is usually found by experiment and has no exact mathematical expression. It is simpler, therefore, to use the following expression:

$$\Delta f = \Delta f_{\mathrm{i}} + \zeta\,[\psi\,(\Delta f_{\mathrm{i}})] \tag{6.7}$$

Assume several values of Δf_{i}, find the corresponding values of Δf, and plot the AFC characteristic as follows.

1. Calculate the dependence of the control voltage, V_{c}, supplied by the frequency detector on the deviation of the intermediate frequency from its nominal value, that is,

$$V_{\mathrm{c}} = \psi\,(\Delta f_{\mathrm{i}})$$

2. Calculate or obtain by experiment the dependence of local-oscillator frequency offset, δf_{LO}, on the control voltage, V_{c}, that is,

$$\delta f_{\mathrm{LO}} = \zeta\,(V_{\mathrm{c}})$$

3. Using the relations obtained in (1) and (2) above, calculate and plot the local-oscillator frequency correction term, $\delta' f_{\mathrm{LO}}$, as a function of intermediate-frequency offset Δf_{i}, that is,

$$\delta' f_{\mathrm{LO}} = \zeta\,[\psi\,(\Delta f_{\mathrm{i}})]$$

Typical plots for ψ and ζ are shown in Fig. 6.16. The arrows in the figure suggest the calculation procedure: entering the plot with Δf_{i}, find V_{c}; entering the next plot with the V_{c} thus found, determine $\delta' f_{\text{LO}}$; use the values of Δf_{i} and $\delta' f_{\text{LO}}$ thus found, locate the points of the curve relating $\delta' f_{\text{LO}}$ to Δf_{i}.

4. Using Eq. (6.7), calculate and plot a curve relating Δf to Δf_{i} (Fig. 6.17*a*). To the dashed line which runs at an angle of 45° when the same scale used on the two axes of coordinates and represents the relation between Δf and Δf_{i} with the AGC loop disabled ($\Delta f = = \Delta f_{\text{i}}$) is added the full line representing the function $\zeta\,[\psi\,(\Delta f_{\text{i}})]$. This characteristic can readily be transformed to a plot of Δf_{i} as a

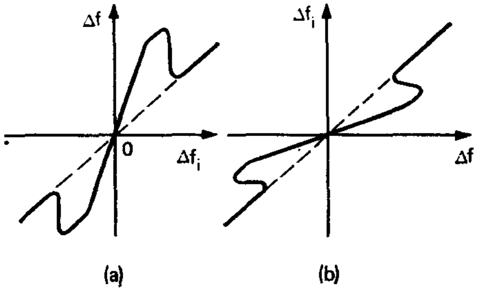


Fig. 6.17

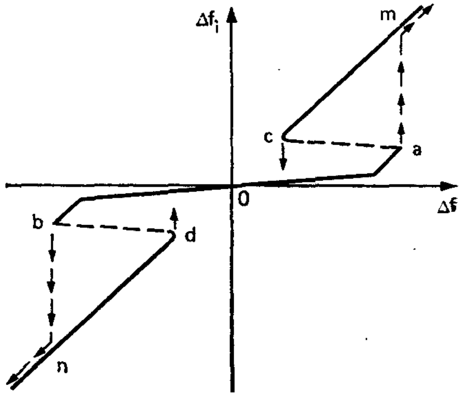


Fig. 6.18

function of Δf. To do this, lay off the values of Δf_{i} on the axis of ordinates, and the respective values of Δf, on the axis of abscissas.

For a properly designed AGC loop, the characteristic looks like that shown in Fig. 6.18. As Δf increases in value, the intermediate frequency deviates from its nominal value increasingly more. Yet, this change in the i.f. is substantially smaller (by a factor of as high as several tens) than the local-oscillator or signal frequency offset that has caused it. This goes on as far as point *a* in the region of positive Δf values and as far as point *b* in the region of negative Δf values. Beyond those points, the characteristic changes into portions *ac* and *bd* shown dashed; they are unstable or, which is the same, do not represent the real behaviour of the process.

The increase in Δf_{i} beyond the portions marked by points *a* and *b* indicates that it has moved outside the operating range of the frequency-detector characteristic, $V_{\text{c}} = \psi\,(\Delta f_{\text{i}})$, as shown in Fig. 6.16. This increase in Δf_{i} is associated with a decrease in V_{c} and, as a consequence, a decrease in the local-oscillator frequency correction term, $\delta' f_{\text{LO}}$. The decrease in $\delta' f_{\text{LO}}$ causes the deviation, Δf_{LO}, of the local oscillator from its nominal frequency to increase, and so does the i.f. deviation, Δf_{i}. This leads to a substantial fall in V_{c}, and the chain of events repeats itself until f_{i} moves well beyond the bandwidth of the i.f. amplifier, and the control voltage supplied by the frequency detector falls practically to zero. As this happens, the device changes stepwise to a new state. The intermediate frequency now takes on the value which it would have had in the absence of automatic frequency control, that is, a value changed by an amount equal to the initial frequency offset, Δf. These events are illustrated in Fig. 6.18: the frequency offset changes abruptly and stepwise from point *a* to point *m* or from point *b* to point *n* (lying on the 45-deg line corresponding to the absence of the corrective action). As Δf increases still more, the representative point keeps moving along the stable portion of the characteristic upwards to the right from point *m* and downwards to the left of point *n*.

Now suppose that the representative point on the characteristic of Fig. 6.18 lies to the right of point *m*, which means that the intermediate frequency lies far outside the characteristic of the frequency detector (see Fig. 6.16). Most often this implies that the intermediate frequency lies outside the passband of the i.f. amplifier, that is, the receiver is completely 'off-tune', and there is no signal voltage present at its input. Now, no control voltage will be applied to the AFC loop.

Let us now decrease Δf, that is, tune the receiver to the desired signal frequency. In consequence, f_1 will decrease, tending towards its nominal value. This will go on until the signal frequency has reached the edge of the passband of the i.f. amplifier and until at least a small voltage appears at the output of the frequency detector (points *c* and *d*). This voltage will servo the local oscillator so as to minimize the deviation of the i.f. from its nominal value. This will in turn build up the control voltage, V_C, and cause the local-oscillator frequency correction term to increase so that f_1 will keep being pulled within the passband of the i.f. amplifier. In consequence, point *c* (and, accordingly, point *d*) is unstable, and it is there that a jump takes place to the stable portion, *aob*, of the AFC characteristic, as shown by the arrows at points *c* and *d*.

As is seen from the plot of Fig. 6.18, a receiver incorporating an AFC loop has two characteristic frequency offset ranges. One of them extends between point *a* and *b*, where the AFC loop 'holds' the i.f. close to its nominal value. Quite aptly, this range is called the *hold-in range.* The other range extends between points *c* and *d* and lies close to the passband of the receiver. When the i.f. falls within this range, the AFC loop 'captures' or 'acquires' the receiver frequency. Following that, the receiver is held 'on-tune', provided the signal frequency swings within the hold-in range. For this reason, the range between points *c* and *d* is called the *capture* or *acquisition range.*

An attempt to tune a receiver having an efficient AFC loop from one station to another may run into a problem. The point is that the AFC loop may lock the receiver to the frequency of the previous, and the operator will not thus be able to tune in the new station. To avoid this, it is practiced to disable the AFC loop during a frequency change by opening or shorting the circuit which feeds the control voltage, V_c.

An electronic AFC loop may show an unstable performance due to a temporary break in signal transmission or signal fading, that is, the variation of radio field strength caused by a gradual change in the transmission medium. In the case of fading, V_c decreases or disappears altogether. At the same time, the local-oscillator frequency undergoes a change because it is a function of V_c: at $V_c = 0$ it takes on the value that it would have in the absence of AFC. If, as a result of this occurrence, the frequency should correspond to some point within portion *cm* or *dn* on the control characteristic (see Fig. 6.18), the exact tuning will not be restored after the signal re-appears at the receiver input. For reception to be resumed, one will then have to tune the receiver by hand so that the local-oscillator frequency falls within the interval *cd*. After that, the AFC loop will acquire the desired frequency and provide for normal signal reception. As an alternative, the receiver can be brought back to tune automatically, but this will call for provision of an additional automatic search-tuning or frequency-sweeping device.

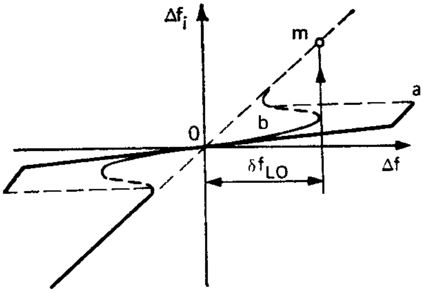


Fig. 6.19

Reception can be upset by partial as well as by complete fading. This is illustrated in Fig. 6.19 which gives AFC characteristics for (a) a normal input signal strength and (b) a reduced input signal strength. It is assumed that the detuning, δf_{LO}, has a value lying anywhere midway between the limits of the hold-in range. Under normal conditions, the i.f. offset is represented by curve *a* and is seen to be small. Therefore, a normal signal reception is provided.

Should, as a result of signal fading, the control voltage V_c fall so that the AFC characteristic takes the shape of curve *b*, the representative point will move during a fade to the edge of the hold-in range of the new characteristic. This is an unstable condition, and the frequency undergoes an abrupt change as shown by the arrow in the figure. The frequency moves outside the passband of the receiver, and no reception of the desired signal is now possible.

As the signal regains its normal strength, the AFC characteristic recovers its original shape, *a*, but the representative point remains in its uppermost position, *m*. In other words, reception will not be resumed, and the local oscillator will have to be tuned anew so that the i.f. falls within the acquisition range.

The above effect introduces an element of unreliability in the operation of a receiver with AFC. It can, however, be avoided if the voltage at the output of the frequency detector has no time to change substantially during a fade. This requirement can be satisfied by increasing the discharge time of the capacitor which is part of the low-pass filter (see Fig. 6.12). Unfortunately, a filter with a long time constant makes the AFC loop insensitive to fast changes in frequency. It is therefore more attractive to enhance the tuning stability of a receiver in any one of the following ways.

1) By changing the time constant automatically through switch-selection of the low-pass filter elements in the AGC loop. When the input signal has a normal amplitude, a short time constant is chosen so that the AFC loop responds to changes in frequency rapidly. As the signal amplitude decreases, a longer time constant is chosen, and V_c slowly decreases for some time. The filter elements can be switched, using the rectified voltage from the output of the i.f. amplifier, for example, from the AGC loop.

2) By using automatic search or frequency sweeping. As the signal fades, a device is enabled to automatically feed a voltage to the control circuit (at *CC* in Fig. 6.15*a*) of the local oscillator. This voltage varies periodically, thus causing the local oscillator to vary or sweep in frequency within a range which encompasses the acquisition range of the AFC loop. After the input signal strength has been restored and the correct frequency has been acquired, the automatic search is discontinued, and normal reception is resumed.

3) By configuring the AFC loop and the receiver as a whole so that the acquisition and hold-in ranges differ only slightly from each other. As has been noted earlier, the acquisition range roughly corresponds to the passband of the receiver, including its frequency detector. Therefore, any spreading in the acquisition range is inevitably associated with the spreading of the overall passband and, in consequence, with an impairment in selectivity. This will not happen if the AFC loop uses a separate, parallel amplifier with an extended bandwidth, or if the passband is reduced as may be required in the succeeding stages of the i.f. amplifier (to the right of point *A* in Fig. 6.15*a*).

6.9. The Frequency Correction Factor

In a receiver with AFC, the i.f. is maintained close to its nominal value. To achieve this, use is madė of the middle, nearly rectangular portion of the frequency-detector characteristic. The small curvature of this portion may play a role in the detection of FM signals because it is responsible for the oocurrence of nonlinear distortion in the received message, but it does not practically affect the operation of the AFC loop. If the frequency does not go beyond the interval *aOb* in Fig. 6.20*a*, it may be taken that

$$V_c \approx S_d \Delta f_1$$

where S_d is the slope of the frequency-detector characteristic. In this case, the steady-state gain of the low-pass filter is either taken equal to unity or lumped with S_d.

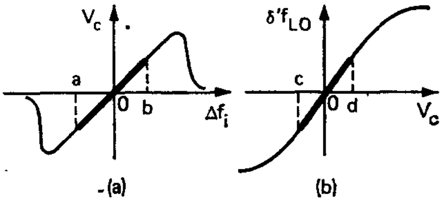


Fig. 6.20

For small i.f. offsets, the voltage V_c applied to the local-oscillator control circuit is low and the local-oscillator frequency is caused to change practically in proportion to this voltage. Therefore, within the portion *cOd* of the characteristic in Fig. 6.20*b*, which is almost linear, it may be taken that

$$\delta' f_{LO} = S_{cc} V_c$$

where S_{cc} is the slope of the control-circuit characteristic. Thus, for small frequency offsets, Eq. (6.7) takes the form

$$\Delta f = \Delta f_1 + S_d S_{cc} \Delta f_1$$

Hence,

$$\Delta f_1 = \Delta f / (1 + S_{cc} S_d) \tag{6.8}$$

or, in a different way,

$$\Delta f_1 = \Delta f / K_{AFC}$$

where

$$K_{AFC} = 1 + S_{cc} S_d$$

is the frequency correction factor characterizing the efficacy of the AFC loop.

6.10. Transients in Automatic Frequency Control

The steady-state frequency offset, Δf_1, as defined by Eq. (6.8), makes any sense only if the AFC loop operates in a stable fashion, without sustained hunting which may, in principle, occur in the AFC loop as in any circuit with feedback.

The control circuit has a practically instantaneous effect on the local-oscillator frequency and may therefore be deemed to be free from time lag or inertia. Transients occurring in the output circuit of the frequency detector can be more significant, but even they play a relatively minor role. As is the case with the AGC circuit, transients occurring in a receiver with AFC in the wake of abrupt changes in the input-signal or local-oscillator frequency depend primarily on the properties of the low-pass filter (*LPF* in Fig. 6.15*a*). To a first approximation, transients in resonant circuits may be neglected for the same reasons as apply in an analysis of transients in a receiver with AGC: the filter should suppress signal modulation products whereas the i.f. amplifier must pass the modulated signal without excessive distortion. In consequence, the low-pass filter in an AFC loop is of necessity the most sluggish element.

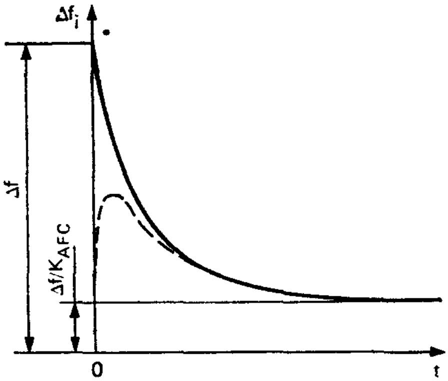


Fig. 6.21

Transients in a low-pass filter may be analysed similarly to transients in an AGC circuit. For relatively small frequency offsets, the control circuit may be assumed to be linear (as for small variations in voltage in Sec. 6.5). The control voltage acting via the control circuit (*CC* in Fig. 6.15*a*) may be written as

$$V_c(p) = K_f(p)\, S_d\, \Delta f_i$$

Accordingly, Eq. (6.8) may be used to analyse transients due to an abrupt change in frequency, Δf, on re-casting it in operational notation as

$$\Delta f_i(t) \to \Delta f(p)/[1 + S_{cc} S_d K_f(p)]$$

or on inserting the expression for $K_f(p)$ in the same form as in Sec. 6.6:

$$f_i(t) \to (1/K_{AFC})\, \Delta f(p) \left[\left(1 + \sum_{k=1}^{n} a_k p^k\right) \Big/ \left(1 + \sum_{k=1}^{n} a_k p^k / K_{AFC}\right)\right]$$

The above expression is not unlike the operational equation for transients in an AGC loop. It follows then that its analysis should yield similar results as well. Notably, for a single-section filter made up of a resistor R_1 and a capacitor C_1 (see Fig. 6.12), the time constant of the control loop is $C_1 R_1 / K_{AFC}$. A salient feature of this case is that the frequency correction factor K_{AFC} may be rather high, running into several tens.

The difference between the transients in the AFC loop and the filter can be explained in the same way as in the case of AGC. Just as the frequency changes by Δf, the voltage at the output of the low-pass filter does not change at once, and the AFC loop remains inactive. For this reason the output frequency of the receiver changes stepwise by Δf, as shown in Fig. 6.21. As the control voltage, V_c, builds up, the i.f. offset Δf_i decreases and approaches its steady-state value, $\Delta f/K_{AFC}$. The frequency-detector output voltage applied to the input of the low-pass filter tracks this deviation, Δf_i, as shown by the full line in Fig. 6.21. The large initial spike in this voltage speeds up the change in the output voltage of the filter and, in consequence, causes the frequency to settle at its steady-state value.

The actual transient response may differ from the chain of events described above for the following reasons.

(1) Owing to transients in the i.f. amplifier, an abrupt change in the signal frequency at the input to this amplifier does not lead to a similarly abrupt change in its output frequency; rather, it changes gradually. Accordingly, the initial portion of the response is smoothened as well, as shown by the dashed curve in Fig. 6.21.

(2) The initial abrupt change in the i.f. frequency, Δf_1, may shoot beyond the portion *ab* of the characteristic, which is, as can be seen from Fig. 6.20*a*, almost linear. Because of this, V_C changes initially by a smaller amount than has been expected, and the effect of the AFC is lowered; therefore, the frequency offset within the initial portion of the response shown in Fig. 6.21 increases.

(3) If the initial abrupt change in frequency goes beyond the receiver passband, the frequency-detector output voltage may fail to provide for frequency acquisition altogether, that is, the representative point will find itself within the portion *cm* or *dn* of the control characteristic shown in Fig. 6.18. For the frequency to be brought back within the receiver passband, one will have to search for the signal automatically or manually.

When use is made of a double-section filter, the condition for an aperiodic transient response may, by analogy with the condition derived for AGC (see Sec. 6.6), be written as

$$K_{AFC} < (1/4)\,[(\tau_1/\tau_2)^{1/2} + (\tau_2/\tau_1)^{1/2} + (C_2/C_1)\,(\tau_1/\tau_2)^{1/2}]^2$$

or

$$K_{AFC} < [(1/4)\,(\tau_1/\tau_2)^{1/2} + (\tau_2/\tau_1)^{1/2} + (R_1/R_2)\,(\tau_2/\tau_1)^{1/2}]^2$$

It is seen that in order to avoid an oscillatory response from an AFC loop, it is essential that the time constants of the filter sections should be sufficiently different in value and that K_{AFC} should not be too great. Transients in the i.f. amplifier may promote the conversion of an aperiodic response of the AFC loop into an undamped (or oscillatory) response. In such a case, no reception of the transmitted information may prove feasible.

With a triple-section filter, the likelihood of an undamped oscillatory AFC response increases to the extreme, which is why AFC loops never use low-pass filters with more than two sections.

6.11. Phase-Locked Loops

The arrangement in Fig. 6.15*b* will operate as what is called a *phase-locked loop* (PLL) if the comparator, *Comp*, is a phase detector. Then, V_C will be a function of the phase difference φ between the voltages of the reference oscillator, *RO*, and of the voltage-controlled oscillator, *VCO*. The PLL like this has a periodic response such as shown in Fig. 5.21 and may be described by the function

$$V_c = \psi(\varphi) K_f$$

where K_f is the transmission gain of the low-pass filter.

Applied to the control circuit, V_c causes the VCO frequency to change by

$$\delta' f = \zeta(V_c)$$

Let the initial frequency offset between the VCO and the reference oscillator be δf. The PLL action will cause the VCO frequency to change by $\delta' f$ so that the frequency offset will be equal to

$$\Delta f = \delta f - \delta' f;$$

As follows from Fig. 5.21, the control voltage V_c may take on a positive or a negative value, depending on the sign of the phase shift φ. In consequence, $\delta' f$ may likewise be positive or negative. In view of the notation adopted above,

$$\Delta f = \delta f - \zeta[\psi(\varphi) K_f] \qquad (6.9)$$

To elucidate the general pattern of events taking place in such a circuit, consider the particular case of a small frequency offset and assume that the portion of the $\zeta(V_c)$ characteristic we are going to use is linear, that is,

$$\delta' f \approx (1/2\pi) S_{cc} V_c$$

Suppose also that there is no low-pass filter, which means that $K_f = 1$. Note furthermore that changes in angular frequency and phase are related as

$$\Delta\omega = 2\pi\Delta f = d\varphi/dt$$

Therefore, Eq. (6.9) may be re-cast as

$$d\varphi/dt \approx \delta\omega - S_{cc}\psi(\varphi) \qquad (6.10)$$

Proceeding from Eq. (6.10), we are in a position to construct a phase portrait and glean an idea about the pattern of variations in the phase of the VCO. Considering the form of the function $\psi(\varphi)$, the phase trajectories for four values of $\delta\omega$ take the shape shown in Fig. 6.22. The arrows mark the fact that when the derivative $d\varphi/dt$ is possitive, the phase angle changes in the positive direction (the representative point moves to the right); when the derivative is negative, the phase angle changes in the negative direction (the representative point moves to the left).

Phase trajectory *a* applies when the VCO has no frequency offset ($\delta\omega = 0$), and shows that the phase shift tends to a steady-state value, $\varphi = 0$. Phase trajectory *b* applies when the VCO frequency has deviated from its assigned value by $\delta\omega$. Here, too, the phase shift tends to a steady-state value but other then zero. In a steady state, the frequencies of the two oscillators are the same, and a constant phase shift, φ_{ss}, is established between the VCO and RO signals. This shift increases with increasing initial frequency offset $\delta\omega$.

If the initial frequency offset, δf, exceeds the maximum value of the correction term, $\delta' f_{\max}$, that can be provided by the phase-detector voltage, $V_{c,\max}$, the trajectory will not cut the axis of abscissas (curve *c*). The sign of the derivative $d\varphi/dt$ will remain unchanged, and the loop will fail to reach a steady state, which means that the PLL fails to servo the VCO to the correct frequency. Hence it follows that the maximum frequency offset that can be handled by a PLL is

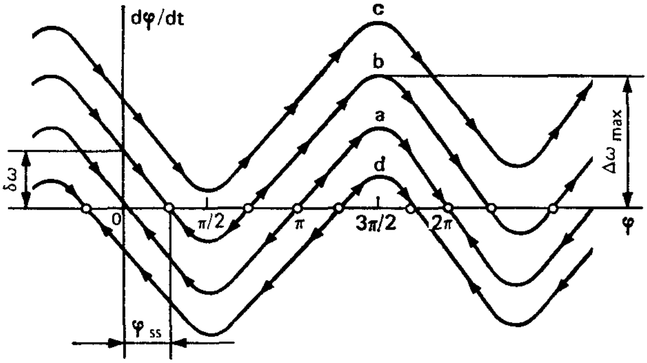


Fig. 6.22

$$\delta f_{\max} = \zeta (V_{c,\max})$$

If a PLL gives the same amount of action with control voltages of both polarities, its hold-in range is $2\delta f_{\max}$. In our case, the hold-in range is the same as the acquisition range because at $\delta\omega < \Delta\omega_{\max}$ the curve cuts the axis of abscissas at two points one of which is stable. If the frequency offset takes the opposite sign (in which case, phase trajectory *d* applies), similar events will occur, but the steady-state position will be shifted relative to zero (or 2π, 4π, etc.) to the left rather than the right.

It is seen from Fig. 6.22 that a PLL may act as a frequency detector. To demonstrate, the steady-state phase shift φ_{ss} depends on or, with the signal frequency only slightly deviating from its assigned value, varies practically in direct proportion to variations in frequency. On the other hand, with small values of the phase shift the phase-detector output voltage is proportional to φ_{ss}. In consequence, it is a linear function, within certain limits, of variations in the frequency of any one of the two oscillators.

Transients in a PLL are difficult to analyse because the loop is nonlinear, and also because they can only be described by high-order differential equations. If we assume that there is no low-pass filter, that the phase detector is free from time lag, and that we may neglect the sluggishness of all other receiver sections, then for small frequency and phase offsets and within the linear portions of the phase-detector and control-circuit characteristics the transients in a PLL may be described by re-casting Eq. (6.10) as

$$\Delta\omega = d\varphi/dt = \delta\omega - S_{cc}S_dK_f\varphi$$

Since,

$$\varphi = \int_0^t \Delta\omega\,(t)\,dt$$

the above equation may be written in operational notation as

$$\Delta\omega\,(p) = \delta\omega\,(p) - S_{cc}S_dK_f\,(p)\,[\Delta\omega\,(p)/p]$$

Hence

$$\Delta\omega\,(t) \to \delta\omega\,(p)\,p/[p + S_{cc}S_dK_f\,(p)]$$

Since t tending to infinity is the same as p tending to zero the above equation bears out that with time the frequency offset approaches zero.

For a single-section low-pass filter,

$$K_f\,(p) = 1/(1 + p\tau)$$

(see Sec. 6.6), and

$$\Delta\omega\,(t) \to \delta\omega\,(p)\,p\,(1 + p\tau)/[p\,(1 + p\tau) + S_{cc}S_d]$$

It is seen that in contrast to an AFC loop, a PLL using a single-section low-pass filter has a second-order characteristic equation

$$p^2\tau + p + S_{cc}S_d = 0$$

whose roots are

$$p = \frac{-1 \pm (1 - 4S_{cc}S_d)^{1/2}\,\tau}{2\tau}$$

Since the real part of the roots is negative, the loop is stable, which means that with time the frequency approaches a constant value. Since, however, $4S_{cc}S_d\tau > 1$, the PLL has an oscillatory response.

With slow variations in frequency, a PLL will ensure a practically perfect synchronization of the two oscillators. Under dynamic conditions, with the frequency varying all the time, the phase shift between the VCO and RO signals will be anything but constant, and there will be differences in frequency, as well.

In a lag-free loop, that is, one using no low-pass filter, and with all the other circuits having a relatively large bandwidth, frequency acquisition would occur instantaneously. As has been noted, the acquisition range would then be equal to the hold-in range. In actual conditions, the acquisition range is always smaller than the hold-in range.

6.12. Automatic Search Tuning

State-of-the-art communications equipment has provisions for exact tuning to any desired frequency and high frequency stability so that there is no need for an automatic signal search if the signal frequency is known in advance. However, a receiver will need search-tuning if the exact settings of the tuning controls are not known in advance.

Search-tuning schemes are many and diverse. A typical arrangement is shown in Fig. 6.23. In this case, the control voltage is supplied by an integrator, *Int*. In turn, the integrator is energized by a voltage which comes from a three-input summator, Σ. Input *1* receives via an electronic switch, *ES*, a direct voltage from a voltage source, *VS*, which produces a slowly rising, or ramp, control voltage, V_c, at the integrator output.

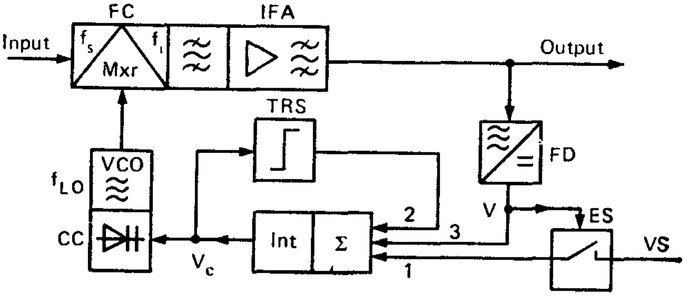


Fig. 6.23

The control voltage, V_c, drives the control circuit, *CC*, so that the frequency of the voltage-controlled oscillator, *VCO*, is varied at a constant rate. At the same time, V_c is applied to a threshold reset circuit, *TRC*. As V_c reaches a predetermined threshold value, the reset circuit feeds to input *2* of the integrator a voltage which has a magnitude and a polarity such that the integrator output voltage is rapidly reset to its initial value, following which the next search cycle begins: V_c rises again until it reaches the predetermined threshold value, and so on. It is easy to see that the integrator and the reset circuit make up between them a sawtooth voltage generator.

Suppose that the mixer, *Mxr*, receives a sufficiently strong signal at frequency f_s. At the instant when the voltage-controlled local-oscillator frequency, f_{LO}, becomes equal to $f_s + f_i$, the difference frequency $f_{LO} - f_s$ becomes equal to f_i, and the signal falls within the passband of the i.f. amplifier. The voltage developed at the i.f. amplifier output is applied to a frequency detector, *FD*. The detected voltage is then applied to input *3* of the summator. At the same time, it operates the electronic switch, *ES*, to break the circuit which supplies the direct voltage from the voltage source, *VS*. From this instant on, the rise in V_c caused by integration of the voltage from the VS is discontinued.

Any further change in voltage depends on the polarity at input *3* of the integrator. Suppose that the i.f. lies below its nominal value and corresponds to point *A* on the frequency-detector characteristic in Fig. 6.24. Now the voltage supplied by the frequency detector is positive. Its integration causes V_c to change, thus leading to a change in the VCO frequency. The voltage is applied in such a polarity that the change in frequency it causes results in a rise in the intermediate frequency (as shown by the arrow in Fig. 6.24) so that it moves closer to its exact value $f_{i,0}$ which is the same as the zero-crossing frequency, $f_{z.c.}$, of the frequency detector. When the two frequencies match, the frequency-detector voltage *V* will be zero, and the VCO frequency will remain unchanged.

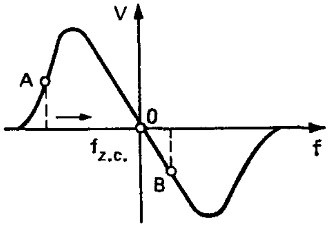


Fig. 6.24

Should the actual i.f. be higher than $f_{z.c.}$ so that it corresponds to point *B* on the characteristic in Fig. 6.24, the integrator will be fed a voltage in reverse polarity. As a result of integration, V_c and the i.f. will change in the reverse direction. In this way, the device continuously servos the VCO (the local oscillator of the receiver) automatically to the required frequency.

In the case of a fade or a break in signal transmission, no voltage will be applied to the integrator inputs, and the integrator output voltages will remain practically unchanged for some time. If the change ('drift') in the integrator output voltage is slow, the associated change in the VCO (local-oscillator) frequency away from the required value in the meantime will be small and will not cause a break in reception. The frequency control system will resume operation as soon as the signal re-appears.

If it is desired to resume or carry on the search for, say, another station, it is required to close the switch *Sw*.

6.13. Heterodyne (Injection) Frequency Synthesizers

As has been shown in Fig. 6.7, the accuracy and stability of the intermediate frequency, which usually is the difference frequency $f_{LO} - f_s$ or $f_s - f_{LO}$, are primarily determined by the stability of the local oscillator. The stability of transmitters and, in consequence, that of the signal frequency is, as a rule, sufficiently high, except for a few cases, such as those related to the Doppler effect. The most efficacious method of generating stable frequencies is to use a thermostatted reference oscillator. Then practically any frequency can be derived from that of the reference source by a synthesizer.

Most synthesizers operate by digital frequency conversion and are adapted for digital control, in which case the desired signal frequency is selected directly by keying in the respective number in digital form. This principle facilitates remote and automatic control. Digital control commands can be generated by hitting push-buttons or manipulating other forms of manual controls or by a microcomputer in the case of automatic control.

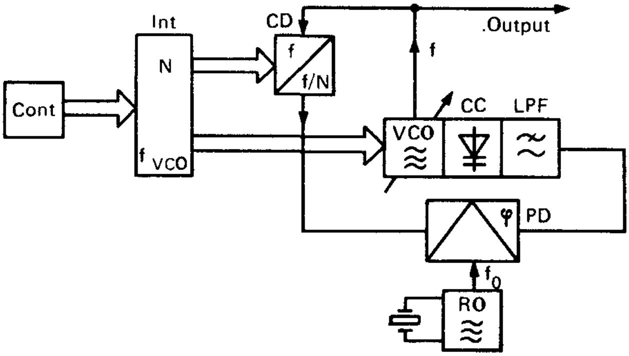


Fig. 6.25

A simplified functional block diagram of a synthesizer which may well serve as a typical subassembly for more elaborate synthesizers is shown in Fig. 6.25. It basically consists of a reference oscillator *RO*, a variable-ratio frequency divider *FD*, a voltage-controlled oscillator *VCO*, and a phase-locked loop.

The variable-ratio frequency divider, *FD*, divides the frequency f of the VCO by N. The resultant voltage at frequency f/N is applied to one of the inputs of a phase detector, *PD*, the other input of which accepts a voltage at frequency f_0 from the reference oscillator, *RO*. The phase-detector output voltage is routed via a low-pass filter, *LPF*, to the control circuit, *CC*, which servos the VCO until the voltages applied to the phase-detector inputs are equal in frequency. This condition answers the equality

$$f/N = f_0$$

Hence,

$$f = f_0N$$

By varying N, it is possible to obtain any discrete frequency increment equal to f_0. This increment may be different, say 10 or 100 Hz for receivers operating in the HF and lower-frequency bands, or 50 kHz or 100 kHz for receivers operating in the VHF and higher-frequency bands. The fractional stability of the synthesized frequency is as good as that of the reference source.

It is not easy to use the most efficient techniques to maintain the frequency stability of oscillators operating at a very low frequency f_0 (for example, 100 or 10 Hz). Therefore, the reference oscillator is usually arranged to generate a higher frequency (most often, 1 MHz) and is crystal-controlled, whereas the desired frequency f_0 is obtained by dividing the reference frequency in a constant ratio.

In order to change the output frequency, the division ratio of the frequency divider should be changed by a manual or automatic control unit, *Cont*. This may, among other things, be a digital programming device (such as a microcomputer) operating via an intermediate control unit usually called an interface, *Int*. In the interface, commands coming from the control unit are converted to signals, and these are applied to the variable-ratio frequency divider so as to set the desired division ratio. If the range of frequency variations of the VCO is greater than what can be handled by the control circuit, the interface tunes the VCO additionally and, if necessary, switches the bands. This action of the interface is labelled in Fig. 6.25 by the arrow drawn from the interface to the VCO.

The very simple synthesizer we have just examined does not provide for frequency variations between arbitrarily broad limits, if the tuned circuit of the VCO has no band selector switch. For this reason it has a limited maximum-to-minimum frequency ratio, K_b. On the other hand, provision of a band selector switch would complicate the synthesizer arrangement and control, and is therefore undesirable. Also, it is to be remembered that the difference between the frequencies at the inputs of the phase detector ought not to extend beyond the acquisition range of the AFC loop. Quite often, one opts for $N = Q + n$, where $n = 0, \ldots, 9$, and $Q \gg n$. Thus arranged, the synthesizer generates output frequencies $f = Qf_0 + \Delta f$, where $\Delta f = nf_0$. If, for instance, $f_0 = 100$ Hz then, by varying n from 0 to 9, it is possible to generate ten output frequencies in steps of 100 Hz within a range of 1 kHz.

The fact that the local oscillator has a low maximum-to-minimum frequency ratio does not mean that the receiver should likewise have a small maximum-to-minimum ratio in terms of the received signal frequency ratio. As has been shown in Sec. 4.5, the infradyne can be tuned between broad limits by varying the frequency of its local oscillator although the latter has a low maximum-to-minimum frequency ratio. Still, the ten output frequencies to which a receiver incorporating the synthesizer of Fig. 6.25 can be tuned are not usually sufficient. Obviously, n should be divided into smaller increments.

Digital frequency synthesis to a specified accuracy is usually done by adding together components of different orders. Each component can be generated by the method illustrated in Fig. 6.25. Then, any output frequency f_s is synthesized so that it answers the formula

$$f_s = 10^m \sum_{1}^{i_{\max}} 10^{i\ \max - i}\, n_i \tag{6.11}$$

where $i = 1, 2, 3, \ldots$; $m = 0, 1, 2, \ldots$; and $n_i = 0$ through 9. If, for example, $i_{\max} = 7$ and $m = 0$, then

$$f_s = 10^6 n_1 + 10^2 n_2 + \ldots + 10 n_6 + n_7$$

Suppose that we have chosen n_i to be $n_1 = 2$, $n_2 = 7$, $n_3 = 4$, $n_4 = 0$, $n_5 = 3$, $n_6 = 7$, and $n_7 = 5$. Then the synthesized frequency will be $f_s = 2\,740\,375$ Hz.

Frequencies of different orders can be combined by a cascade of mixers and a synchronized oscillator. The block diagram of a device implementing this principle is given in Fig. 6.26. The reference source is a PLL-controlled oscillator, *Osc*. The control voltage comes from a phase detector, *PD*, via a low-pass filter. The desired output frequencies are produced by mixers Mxr_1, Mxr_2, etc., with their outputs coupled to a respective filter. Each filter is arranged to pass only the ten frequencies ($n_i = 0$ through 9) that can be supplied by the mixers.

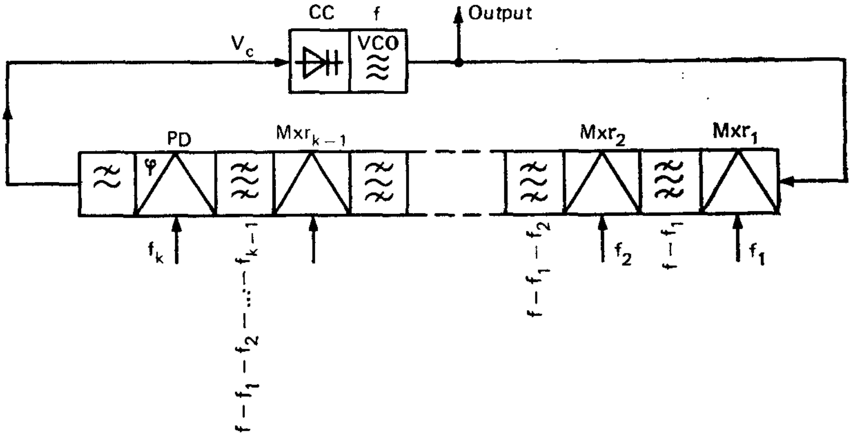


Fig. 6.26

If the VCO frequency is f, then the output frequency of the first mixer, Mxr_1, whose heterodyne input accepts frequency f_1, will be $f - f_1$. On passing through the first filter, $Filt_1$, it is applied to the second mixer, Mxr_2, which at the same time is fed a second frequency, f_2. As a result, the second filter, $Filt_2$, separates a frequency equal to $f - f_1 - f_2$, and so on. The $(k - 1)$st mixer produces a frequency equal to $f - f_1 - f_2 - \ldots - f_{k-1}$ which is applied to one of the inputs of a phase detector, PD, the other input of which accepts the last-combined frequency f_k. The phase-detector output or error voltage, V_c, servos the voltage-controlled oscillator, VCO, to minimize the frequency offset. As a result, the frequencies applied to the phase-detector inputs become exactly equal, that is, $f - f_1 - f_2 - \ldots - f_{k-1} = f_k$. It follows then that the steady-state frequency of the VCO is

$$f = f_1 + f_2 + \ldots + f_{k-1} + f_k$$

which was to be proved.

Since frequencies $f_1, f_2, \ldots, f_k$, each of which can be generated by the technique illustrated in Fig. 6.25, correspond to a particular decimal place in the numerical value of frequency f, each constituent synthesizer should be fed a reference frequency of the respective order. These frequencies can be derived by a cascade of consecutive dividers arranged as shown in Fig. 6.27, where RO is the reference oscillator.

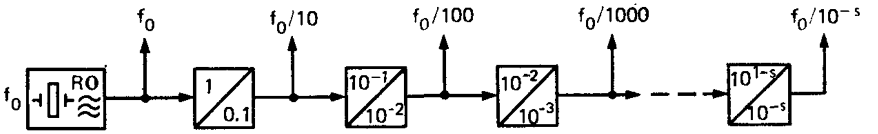


Fig. 6.27

The above principle can be implemented by a synthesizer arranged as shown in Fig. 6.28. In this arrangement, the frequencies are combined as explained earlier and illustrated in Fig. 6.26; each of the combined frequencies is synthesized as illustrated in Fig. 6.25, and the reference frequencies are derived by variable-ratio frequency dividers, as shown in Fig. 6.27. The desired division ratios for the frequency dividers FD_1 through FD_k are set by a control unit, CU.

For further insight into operation of the synthesizer shown in block-diagram form in Fig. 6.28, assume that the division ratio of the second variable-ratio frequency divider, FD_2, is N. Suppose also that the heterodyne input of this divider is fed via fixed-ratio dividers a reference frequency, $f_0/10^s$. Owing to the action of the PLL of the second oscillator, Osc_2, which feeds the second phase detector, PD_2, one obtains

$$f_2/N = f_0/10^s$$

and the second oscillator, Osc_2, is obliged to generate a frequency

$$f_2 = (f_0/10^s)N$$

For the reasons explained earlier, it is assumed that

$$N = Q + n$$

where Q is a constant and n ranges from 0 to 9. The remaining frequencies, that is, $f_1, f_2, \ldots, f_{k-1}, f_k$, are synthesized in a similar way.

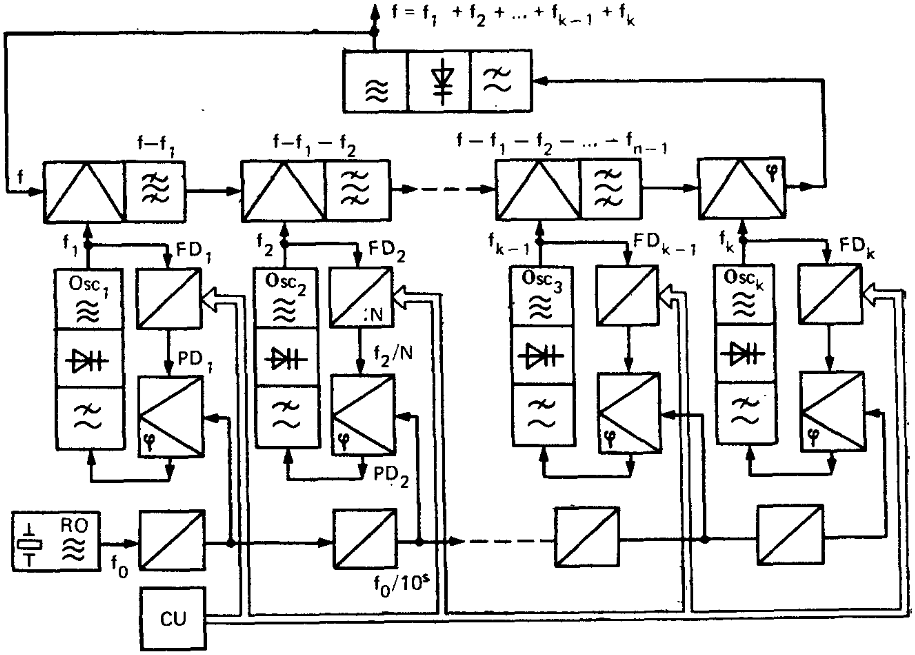


Fig. 6.28

The technique set forth above can produce frequencies given by

$$f = 10^m \sum_{1}^{i_{\max}} 10^{i_{\max}-i}(Q_i + n_i) \tag{6.12}$$

For example, if $m = 0$ and $i_{\max} = 6$, then

$$f = (Q_1 + n_1)10^6 + (Q_2 + n_2)10^5 + (Q_3 + n_3)10^4 \ldots$$

or

$$f = f_\Sigma + n_1 \times 10^6 + n_2 \times 10^5 + n_3 \times 10^4 + \ldots$$

where

$$f_\Sigma = 10^m \sum_{1}^{i_{\max}} 10^{i_{\max}-i} Q_i$$

In order to synthesize frequency f_s answering Eq. (6.11), one needs the difference frequency

$$f_s = f - f_\Sigma$$

Frequency f_Σ is obtained by adding together the fixed frequencies derived from the reference-source frequency by the fixed-ratio dividers with a division ratio $N_i = 10^i$.

In the general case, one can build a synthesizer with any lower frequency of the range, $f_{\min}$. To this end, one should use a mixer in which one input accepts a frequency $f_\Sigma - f_{\min}$. Then the output frequency of the mixer will be

$$f - (f_\Sigma - f_{\min}) = f_{\min} + n_1 \times 10^6 + n_2 \times 10^5 + n_3 \times 10^4 + \ldots$$

6.14. Digital Frequency Indication and Digital Frequency Control

Until quite recently, the tuning indicator was the tuning dial calibrated in units of frequency and/or wavelength. Since there is always a limit to the size of a tuning dial, the reading accuracy was limited as well. A major step forward was made with the advent of digital readout indicators which at first used incandescent lamps and glow-discharge gas tubes. At present, this is done by edge-illuminated panels and liquid crystals. Since they draw power very sparingly, liquid crystals are especially popular in equipment with power supplies of limited capacity.

The basic form of digital frequency indication is the one using a digital pulse counter, that is, a circuit which counts the number of pulses per unit time, converts the count to a frequency, and presents the result in digital format. For this purpose, the voltage whose frequency is to be determined is converted to pulses by the usual methods, such as an arrangement made up of a pulse-height limiter, a differentiator, and a rectifier. The most common digital display in modern equipment is the seven-segment readout. It consists of seven illuminated bars arranged in a figure-of-eight pattern so that all the ten decimal digits can be formed by lighting appropriate bars.

The count in each decimal place (say, tens of megahertz, hundreds of kilohertz, etc.) are converted to binary numbers (four binary digits are enough to form the decimal digits from zero to 9). The binary-coded signal is then fed to a decoder which has seven outputs to the readout segments. If high-power displays are used, such as incandescent lamps or glow-discharge tubes, an amplifier will have to be inserted between the decoder and the readout. In approximate form, the circuit of a one-digit readout is shown in Fig. 6.29. Here, *C* is the counter, *D* is the decoder, *Amp* is the amplifier, *RO* is the readout, and *TB* is the time-base oscillator which maintains the time interval over which the number of signal cycles is counted. For this purpose, the frequency of the crystal-controlled time-base oscillator is divided by frequency dividers; the resultant signals with a period equal to the time base are used to form a start pulse and a stop pulse to start and stop the counter.

In a receiver, there is a special requirement for tuning frequency indication. The point is that whatever the signal frequency f_s, it is above all necessary to tune the local oscillator so that the i.f. falls within the passband of the i.f. amplifier. Because, in the course of tuning, the signal may be weak or absent altogether, it may be difficult, if at all possible to measure the signal frequency. Therefore, instead of the signal frequency, the counter determines the local-oscillator frequency. On the other hand, it must be so arranged that instead of the local-oscillator frequency which is actually being determined the indicator should display the signal frequency to which the receiver is tuned.

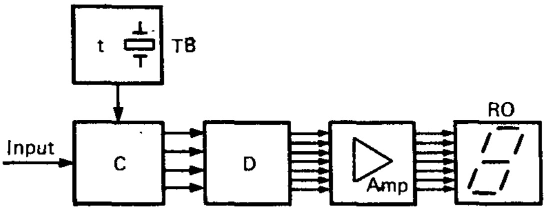


Fig. 6.29

If $f_{LO} = f_s + f_i$ and, in consequence, $f_s = f_{LO} - f_i$, the tuning frequency can be indicated by what may be called the "down-count" technique. The counter is connected to the local oscillator for a time required for the count to represent the local-oscillator frequency at the end of that time interval. This time interval may be taken equal to 1 s, if the accuracy should be within 1 Hz. As an alternative and according to the required accuracy of indication, the count interval may be taken equal to 0.1 s, 0.01 s, etc. The count thus obtained is stored and the number of pulses equal to f_i is then counted so that they are subtracted, or down-counted from the previous result. At the end of the selected time interval, the final count will be $f_{LO} - f_i$, or f_s, which is what is required.

In another, or "up-count", technique, the counter is initially set to read a count which is below zero by the value of f_i. Then, during the selected time interval the number of pulses corresponding to f_{LO} will be counted, or algebraically subtracted from the preset value, and the result actually displayed will be $f_{LO} - f_i$.

Sometimes, a receiver may be extended to include an additional fixed-frequency oscillator operating at f_i. In simplified form, the connection of a tuning indicator is shown in Fig. 6.30. Here, *LO* is the local oscillator used in the frequency converter, *FFO* is the fixed-frequency oscillator operating at f_i, and *TB* is the time-base oscillator which sets the time interval during which the count is accumulated. This time interval is divided into two equal parts. During the first half-interval, switch Sw_1 is closed, and the local-oscillator frequency is counted. During the second half-interval, switch Sw_1 is open, switch Sw_2 is closed, and the *FFO* frequency is counted. By the end of the selected time interval, the counter, *C*, feeds to the readout a result equal to the sum $f_{LO} + f_i$, which will correspond to the signal frequency if $f_i = f_s - f_{LO}$. If, on the other hand, $f_i = f_{LO} - f_s$, a reversible or up-down counter should be used to count pulses in one direction during the first half-interval and in the opposite direction during the second.

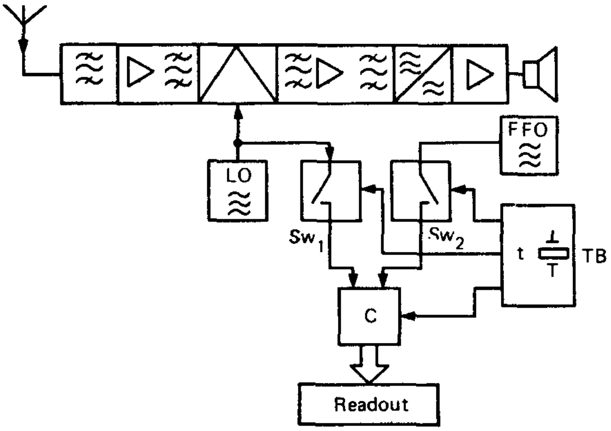


Fig. 6.30

Less accurately but in a far simpler way the local-oscillator frequency f_{LO} can be measured, using the arrangement shown in Fig. 6.31. It operates when there is a signal at the i.f. amplifier output. Instead of a fixed-frequency oscillator operating at f_i, use is made of the signal voltage taken from the output of the i.f. amplifier.

With digital frequency indication, it is possible to build a receiver which has no frequency synthesizer but which shows performance comparable with a receiver having a frequency synthesizer. The task is achieved by tuning the receiver to the required frequency with the aid of a digital readout and a digital AFC for tuning stabilization.

For example, a receiver using digital frequency indication and arranged as shown in Fig. 6.30 can be tuned to the desired frequency with an accuracy determined by the number of digits presented by the readout. With a synthesizer, this would be a highly stable frequency. With a conventional local oscillator, however, there occurs what is known as *frequency drift* caused by variations in ambient temperature and other external factors. The rationale of a digital AFC loop is based on the fact that a change in the numerical value of the frequency always commences at the last (or the least-significant) digit of the number.

The AFC loop is activated each time the least-significant digit changes either way and restores it and, as a consequence, the tuning frequency to the original value. The desired least-significant digit can be written (by, say, hitting a button) into a memory module and later compared (by a digital comparator) with the actually presented least-significant digit. Should the two differ, the AFC loop will go into action.

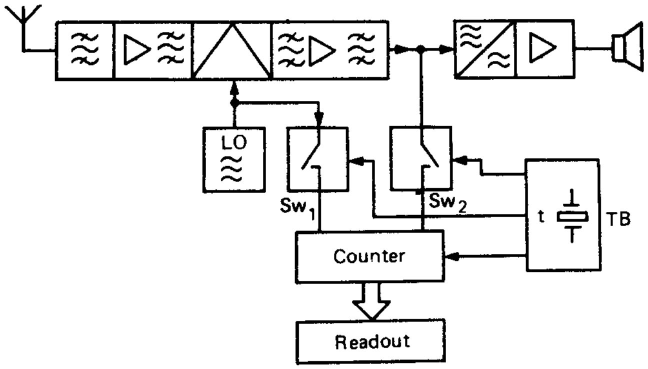


Fig. 6.31

Ordinarily, the memory and the comparator are electronic devices. For simplicity, however, Fig. 6.32 shows them in the form of a simple electromechanical model in which the functions of both are performed by a manually operated rotary switch with semicircular contacts.

The frequency *f* to which the receiver is tuned is displayed by the readout, *RO*. In a binary-coded format (see Fig. 6.29), the least significant digit is transferred to a 10-lead decoder, *D*. After decoding, the pulses are fed into the lead assigned to the least-significant digit. In our example, this is the numeral "6" (see Fig. 6.32). In Fig. 6.32, the readout displays a seven-digit number in which the least-significant digit (for purposes of illustration) is 6.

The desired value of the least-significant digit is set into the memory device by positioning the gap *G* between the semicircular contacts opposite the contact pin connected to the decoder lead associated with the required digit. The semicircular contacts are connected to the inputs of amplifiers Amp_1 and Amp_2, respectively. The inverter provided at the output of the second amplifier reverses the polarity of the pulses. Should the least-significant digit in the readout change, the pulses from the decoder will be conveyed by the lead corresponding to its new value. Passing via the switch, they will arrive at the input of one of the amplifiers in a polarity corresponding to the sign of the deviation. Proceeding further via a low-pass filter, *Filt*, they will drive the control circuit, *CC*, of the local oscillator. In the example of Fig. 6.32, should the digit 6 change to 5 or 4, pulses will go to the first amplifier, Amp_1. Should it change to 7 or 8, they will go to the second amplifier, Amp_2.

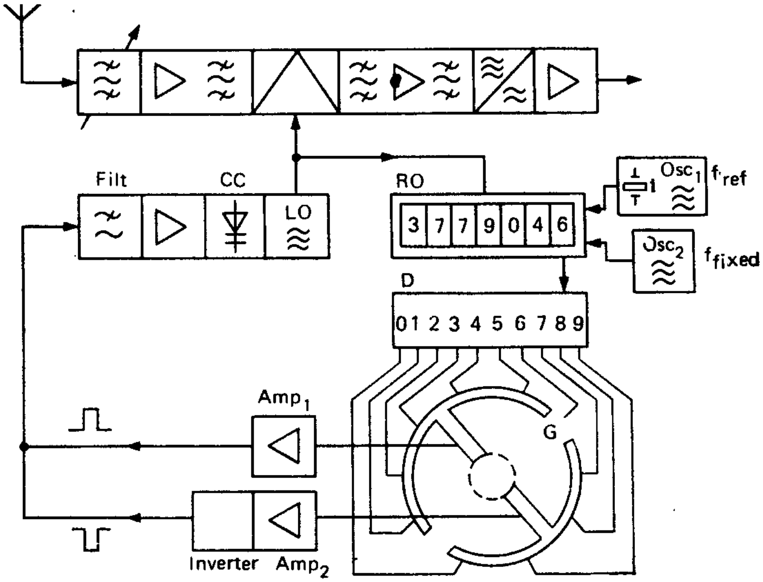


Fig. 6.32

After the AFC loop has restored the correct least-significant digit in the number representing the value of the frequency, (which is 6 in our example), pulses cease to be fed to the AFC loop.

6.15. The Use of Microprocessors in Automatic Tuning

Owing to integrated-circuit technology, practically unlimited number of circuits and circuit components can be built into a tiny chip. No less importantly, the cost of such devices only slightly depends on their complexity. This has removed any psychological and economic restraints for the designer in his quest for ever more sophisticated configurations leading to ever better technical and ergonomic properties of receivers. Special promise is held by digital devices, notably, microprocessors. The potentialities they have brought with them have not yet been fully explored and the effort to put them to use will continue, but some general trends have already taken a sufficiently clear shape.

As a rough idea about the devices providing for control of receiver tuning, Fig. 6.33 shows the most typical of their units. Here, *Rvr* is the receiver, and *CU* is the control unit. It is easy to see that in terms of complexity the control unit is comparable with or is even more complicated then the receiver itself. If it were not microelectronics, the idea to use such devices would have never occurred.

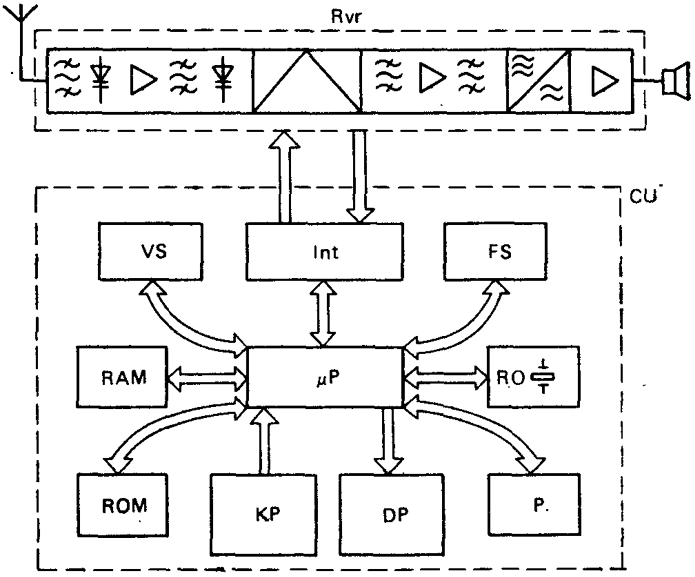


Fig. 6.33

The heart of the control unit is a microprocessor, μ*P*, coupled to the receiver by way of an interface, *Int*. The data stream flowing from the receiver via the interface to the microprocessor carries information about the presence (or otherwise) of the signal from the desired transmitting station, signal strength and other signal properties, coded command signals, and the like. The reverse data flow includes signals that activate the automatic signal search device (see Sec. 6.12) and the AFC loop which mute the a.f. section for the duration of tuning (this is so-called quiet tuning often used in broadcast receivers), and signals to receiver indicators and meters.

The microprocessor is coupled to a reference oscillator, *RO*, which is essential for the functioning of the frequency synthesizer, *FS*, and also generates time marks in response to which the programs stored in the programmer, *P*, are fetched and executed. The voltage required for electronic tuning is supplied by a voltage synthesizer, *VS*. The data which are essential for receiver control and which call for long-time storage (station frequencies, transmission timetables, and the like) are stored in a ROM module. There is also a RAM module to store instructions necessary for operation of the microprocessor.

If and when necessary, the operator can key in certain commands manually with the aid of a keypad, *KP*. The operational status of the receiver, frequency settings, time of day are all presented by luminous readouts arranged on a display panel, *DP*.

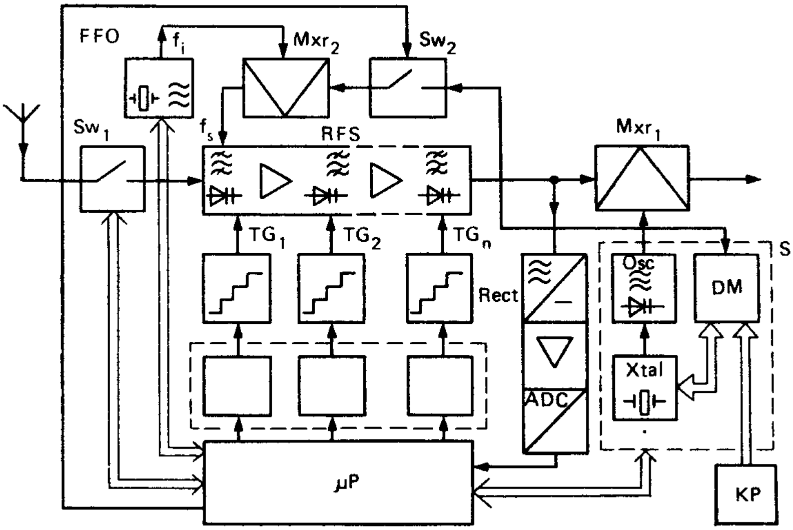


Fig. 6.34

The AGC, AFC and PPL systems we have examined in the earlier sections serve each to control only one variable which may be the gain or the local-oscillator frequency. More elaborate systems, such as shown in Fig. 6.33, equipped with the necessary software, a microprocessor, ROM and RAM modules, have capabilities for multivariable receiver control subject to the requirements that should be met in the face of internal and external destabilizing factors. Instead of a single control algorithm, such system can perform various actions in any preprogrammed sequence.

As an example, consider the case of tuning a receiver to a desired frequency (see also Fig. 4.2). An approximate arrangement of the circuits that control the tuning of the tuned circuits in the receiver is shown in Fig. 6.34. Here, *RFS* is the radio-frequency (r.f.) section.

Using his keypad, *KP*, the operator keys in the desired signal frequency, f_s, to the digital module, *DM*, and the synthesizer, *S*, proceeds to derive the required local-oscillator frequency, f_{LO}, with the aid of an oscillator, *Osc*, from the oscillations of a crystal unit, *Xtal* (see Sec. 6.13). This frequency is then fed to the heterodyne input of a mixer Mxr_1. At the same time, signals from the microprocessor, μP, open switch Sw_1 so as to disconnect the receiver from the antenna, and close switch Sw_2. Via this switch, f_{LO} is applied to the signal input of an auxiliary mixer, Mxr_2. The heterodyne input of Mxr_2 accepts the output of a fixed-frequency oscillator *FFO*, which generates voltage at f_i. The output frequency of Mxr_2 is the signal frequency to be received, that is, $f_s = f_{LO} - f_i$ or $f_s = f_{LO} + f_i$. This frequency can be displayed on a digital readout, and the operator is in a position to monitor the tuning procedure.

The voltage at frequency f_s appearing at the Mxr_2 output is hundreds of times stronger than the signal induced in the antenna. It is routed to the input circuit of the receiver and goes through to the output of the r.f. section, although the latter is not yet tuned. This voltage may be used to tune the tuned circuits.

The microprocessor turns on a generator which produces pulses that go to a digital-to-analog converter serving as a tuning generator, TG_1, which supplies the tuning voltage for the input tuned circuit of the receiver. As the resonant frequency of the tuned circuit approaches f_s, the output voltage of the r.f. section builds up. Its build-up is monitored by a network made up of a rectifier, *Rect*, and an analog-to-digital converter, *ADC*, whose output is coupled to the microprocessor. On passing through its peak, the output voltage of the r.f. section begins to decline; at that instant, incremental (discrete) tuning is stopped, and a command from the microprocessor causes the tuning voltage to vary in the reverse direction one increment or step at a time so as to restore the condition of resonance. Subsequently, the tuning voltage is maintained at the value thus found.

A similar sequence of events applies to a second tuning generator, TG_2, which tunes the tuned circuit in the r.f. amplifier. All the succeeding tuning generators and the associated circuits operate in a similar fashion. After the tuned circuits have all been tuned, switch Sw_2 opens and the voltage at f_s, generated by Mxr_2, is removed from the receiver input. At the same time, switch Sw_1 closes and connects the receiver to the antenna. This completes the tuning procedure.

6.16. Bandwidth Control

In the absence of strong noise and interference, the bandwidth (or passband) of the i.f. amplifier is chosen to be sufficiently wide to accommodate the most essential portion of the r.f. signal spectrum, so that the transmitted intelligence could be reproduced without noticeable distortion. Too narrow a bandwidth would lead to distortion, whereas too wide a bandwidth would impair selectivity.

If a receiver is intended to handle widely varying signals, such as AM or SSB telephone signals, variously keyed or modulated telegraph signals, and the like, it is obvious that such signals will inevitably differ in the width of the frequency spectrum. Quite logically, there will be a need to adjust the receiver bandwidth so that it matches the spectrum width in each particular case. In fact, such bandwidth control will serve a useful purpose even when the signal spectrum is constant but the receiver operates in the presence of strong noise or interference.

In the presence of strong fluctuation noise, the reduction in the receiver bandwidth will bring down the noise level at the receiver output. This band-limiting is accompanied by signal distortion, but the signal level will usually change only slightly. This is the case, notably, with pulse signals. When the bandwidth is wide, the signals experience a minimum of distortion. When the bandwidth is reduced, the signal settling time increases, but it remains shorter than the pulse duration and the signal is able to reach its normal value, whereas the noise level is brought down. With a further reduction in bandwidth, however, so that the settling time (which is inversely proportional to bandwidth) exceeds the signal duration, the signal voltage will have no time to reach its maximum magnitude while the pulse exists, and the pulse height will be reduced appreciably. Obviously, there must be an optimal bandwidth; its value depends on noise strength.

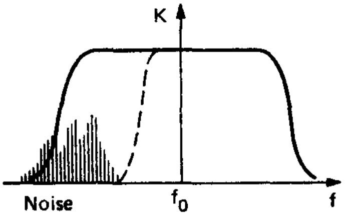


Fig. 6.35

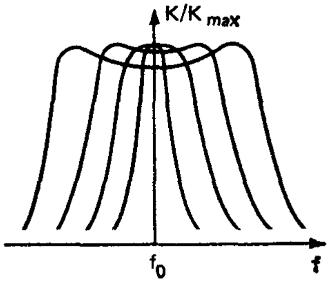


Fig. 6.36

The frequency response of a receiver may be represented by a curve similar to the unbroken curve in the plot of Fig. 6.35. If there is a strong noise present in the adjacent channel (its frequency spectrum is shown at the edge of the bandwidth), it can be reduced by reducing the bandwidth as shown by the dashed curve. As a result, the received message will still be corrupted, but to a lesser extent.

Stepwise bandwidth control can be effected by simply switching the required filters; instead of a filter having a large passband, one can insert a filter having a medium or a narrow passband. The necessary change can be done manually using a mechanical switch, or electronically.

Continuous bandwidth control is the simplest of all to effect by varying the damping factor of, or the amount of coupling between, the tuned circuits that make up the band-pass filter, or by stagger-tuning the tuned circuits. In the former case, the tuned circuit is shunted by a variable-impedance network (see Fig. 6.6). The lower the impedance of the element shunting a tuned circuit, the greater the damping factor of the latter and, as a consequence, the wider the passband. At the same time, however, the rate of rolloff of the frequency response decreases, and its selectivity is impaired.

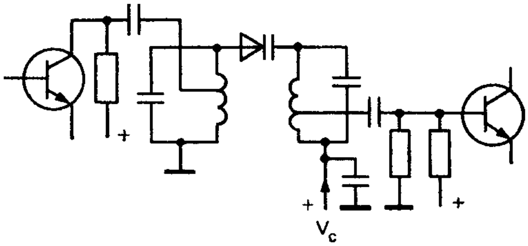


Fig. 6.37

By varying the amount of coupling between the tuned circuits which make up, say, a double-tuned bandpass filter (see Sec. 3.11), it is possible to vary the bandwidth without affecting the rate of rolloff of the frequency response. An approximate form of the frequency response thus obtained is shown in Fig. 6.36. This can best be done with the aid of a varactor, as shown in Fig. 6.37, but there are also other methods.

Bandwidth control by stagger-tuning of resonant circuits is illustrated in Fig. 6.38. If two stages are tuned to the same frequency, the bandwidth is a minimum, as can be seen from Fig. 6.38*a*. When the stages are stagger-tuned, the bandwidth is spread, as shown in Fig. 6.38*b*, but the gain is reduced. This is usually tolerable because the bandwidth is usually spread in the case of strong radio signals when noise and interference can only slightly affect the quality of reception.

Tuned circuits can be stagger-tuned and their frequency can be controlled by AFC, using varactors. The manner in which the required control voltage is derived should be chosen according to the purpose of control. The device supplying this voltage can evaluate the signal-to-noise ratio and control the bandwidth so as to maximize this ratio. The signal spectrum within the bandwidth can be analysed, the signal can be separated, the noise or signal-distortion level can be evaluated, the control voltage can be generated, and the search for an optimal set of operating conditions can be carried out by a microprocessor.

In simplified automatic bandwidth control circuits, the required control voltage comes from the AGC circuit. This voltage is a maximum in the case of a strong signal. It spreads the bandwidth, thus improving the reproduction of the transmitted intelligence without an appreciable increase in the effect of noise because the signal-to-noise ratio is high. With a weak signal, the AGC voltage goes down and the bandwidth is reduced, which is an advantage because noise is felt more with weak signals.

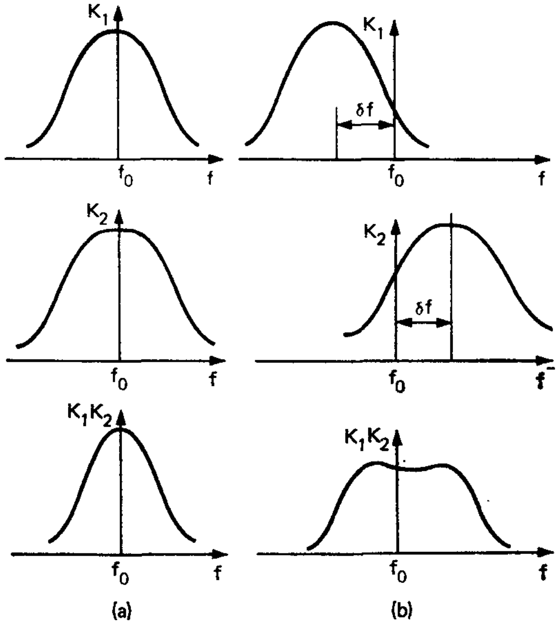


Fig. 6.38

6.17. Touch Controls

The use of controls often involves making, breaking and switching of circuits. As a way of simplifying construction, reducing the overall size, and enhancing speed of response and reliability, it is widely practiced to use switches in the form of electron devices operating in the ON-OFF mode. Apart from being labour-consuming in manufacture and not easy to be miniaturized, electromechanical switches are not sufficiently reliable because of wear and tear, notably that of their contacts. These factors have promoted the advent of what is variously called feather-touch, soft-touch, or, simply, touch controls. With this type of controls, it is enough merely to touch the pad, key, or button bearing an appropriate symbol, letter or numeral in order to switch circuits, to change operating conditions, or perform some other function.

Radio receivers use touch switches which will make or break the associated circuits for as long as the operator holds his finger to the respective control, or effect a circuit-switching action without a reset action after the operator withdraws his finger. In the latter case, he will have to touch the control for a second time in order to restore the previous condition.

In a receiver, the typical functions of touch switches are as follows: turn on the automatic search tuning system, select prese t frequencies, mute the a.f. amplifier while tuning a broadcast receiver (quiet tuning), disable the AFC loop when going from one frequency to another (in this particular case, the touch switch is physically integrated with the manual tuning knob and operates just as the knob is touched), turn on illuminated dials, indicators, and the like for the duration of manual tuning or adjustment.

As a circuit is switched, it is usual for the associated indicator to light up so as to confirm that the command has been executed.

One of the likely arrangements using touch switches with which any one of several preset frequencies can be selected is shown in Fig. 6.39. Here TS_1, TS_2, . . ., TS_n are touch switches each of which is associated with a particular preset frequency; P_1, P_2, ..., P_n are pulsers which supply control pulses and which operate as the respective switch is touched; M_1, M_2, . . ., M_n are flip-flops used as memory units. When a signal is applied to input a_i from the respective pulser, P_i, this causes the associated control-voltage source, CVS_i, to turn on. In the simplest case, this may be a voltage divider, such as shown in Fig. 2.9.

Should input b_i of the same flip-flop, M_i, simultaneously accept a control pulse from another pulser via one of the OR$_i$ circuits shown in Fig. 6.39, no control voltage will be supplied by CVS_i.

The control voltage generated by CVS_i is applied via a common output circuit, OC, which usually is an OR gate, to the varactors in the controlled tuned circuits. Suppose that the receiver has been set to the 2nd preset frequency out of a complement of fixed frequencies by pressure on the touch switch TS_2, and that the control voltage routed via the output circuit, OC, to the varactors in the controlled tuned circuits comes from CVS_2. Now let the operator touch, say, TS_1 which selects the 1st preset frequency. This will cause a control signal to be routed via TS_1, P_1, M_1 to CVS_1, and the latter will supply via the output circuit the control signal corresponding to the new frequency setting. In turn, the output circuit will send it to the varactors in the controlled tuned circuits. At the same time, the control signal will go to the $OR_2, \ldots, OR_n$ gates from which it will be applied to inputs $b_2, \ldots, b_n$ of the flip-flops M_2 through M_n. As a result, the flip-flops will be reset, the signal they have stored will be cleared, the control signal assigned to the 2nd preset frequency will no longer pass through M_2, and CVS_2 will cease to supply a control voltage. In this way, the receiver will be tuned to the 1st preset frequency. Similarly, by touching TS_3 the operator can set the receiver to the 3rd preset frequency, etc.

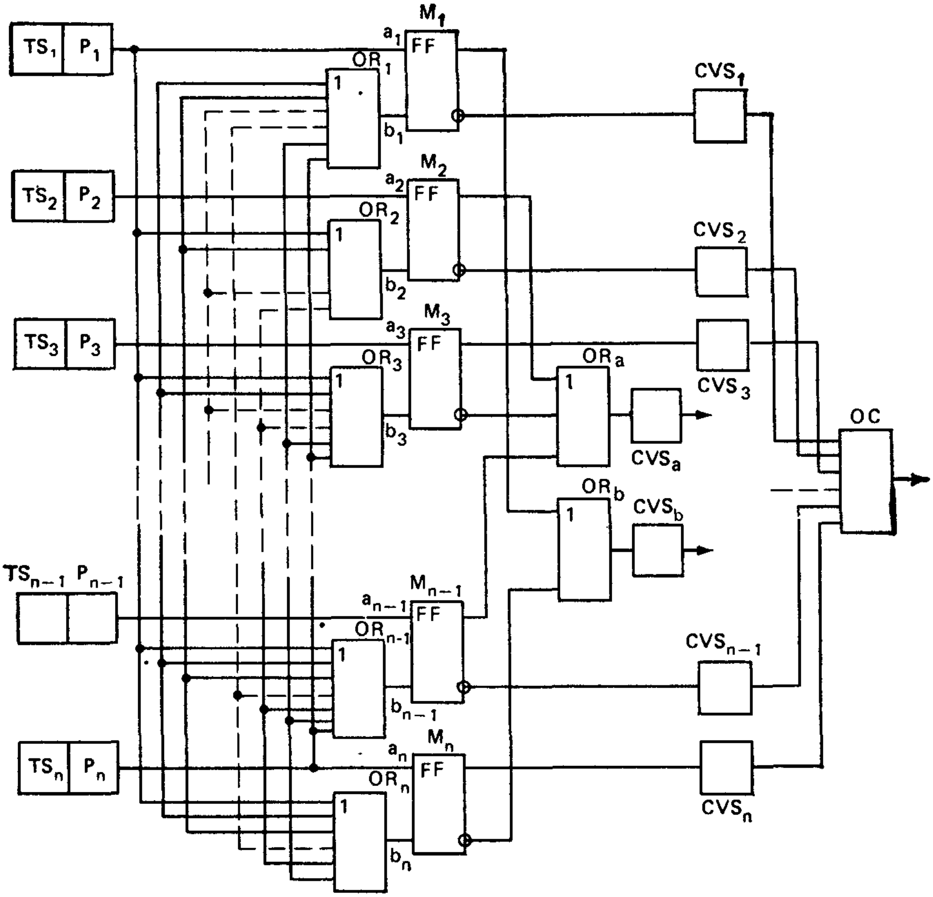


Fig. 6.39

Tuning may involve the switching of other circuits, for example in order to select a particular frequency band, to change over from AM to FM reception, and the like. This job is done by additional circuits likewise controlled by signals from M_1 through M_n in appropriate combinations. As an example, Fig. 6.39 illustrates the case in which additional control signals are obtained from CVS_a and CVS_b; their number may be any. CVS_a is activated by touching TS_2, TS_3 and TS_{n-1}, whereas CVS_a is activated by touching TS_1 and TS_n. The signals are combined by the OR_a and OR_b logic gates.

Touch switches come in many and diverse designs and arrangements; two of them are shown in Fig. 6.40. The touch switch in Fig. 6.40*a* depends for its operation on changes in the circuit resistance as the operator's finger touches the electrode *E* because the skin of the finger has a conductance of tens of microsiemens. This conductance acts to reduce the potential at the base of the input transistor, T_1. This also causes a reduction in the emitter potential, with the result that the Schmitt trigger built around T_2 and T_3 changes state.

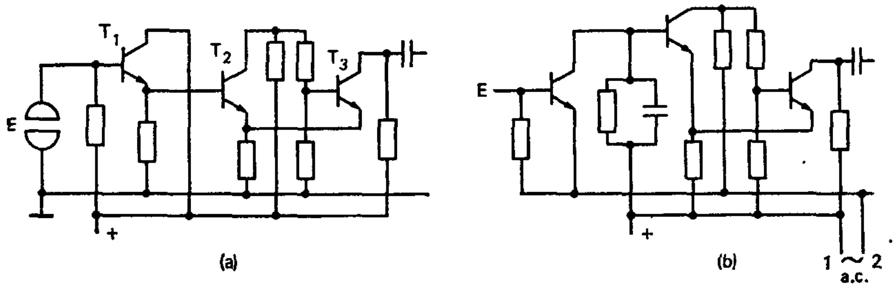


Fig. 6.40

The touch switch shown in Fig. 6.40*b* depends for its operation on changes in the alternating voltage induced at the input electrode of the transistor via the conductance associated with the operator's body. The alternating voltage applied between terminals *1* and *2* may come from a supply line or any other source. As the electrode *E* is touched, which is equivalent to placing the human body's resistance between the base of T_1 and ground, an alternating voltage is produced between the emitter and base of the input transistor. It is amplified, rectified and applied to a Schmitt trigger, thus causing it to change state. In each case, the change of state by the Schmitt trigger causes an output voltage pulse to be generated.

Chapter Seven

Radio Noise and Interference and Their Suppression

7.1. Classification and General Characterization

The medium through which radio waves are transmitted contains discontinuities which are responsible for wave-energy absorption and scattering with time *t*, frequency *f*, and along spatial, *r*, and polarization, p, coordinates. The result is what is called multiplicative noise responsible for random signal distortion which shows up as fading, multipath interference, pulse edge distortion, and intersymbol interference. The transmission medium, or the channel, can best be described by the transmission tensor $\mathbf{H}$ (t, f, $\mathbf{r}$, p). Unfortunately, this tensor is difficult to use because one then has to consider a host of factors, their interrelations, and lack of statistical data. That is why it is customary to characterize the channel by giving its transfer function μ (t) which is a special case of the tensor $\mathbf{H}$ ($\cdot$), given some fixed values of f, $\mathbf{r}$, and p.

In the general case, the transmission medium can be treated as a stationary, linear system with random variables. However, the physical nature of real channels is such that there exist local time slots T_{st} inside which the received signals and noise may be treated as stationary, random processes. According to experimental data, for ionospheric channels in the HF band and for tropospheric channels T_{st} ranges from 5 to 15 min.

Apart from multiplicative noise, the signal during its travel in the channel is acted upon by additive noise produced by various sources of electromagnetic radiation.

The term "radio noise" refers to any electromagnetic disturbances which fall within the r.f. range and impair the quality of message reception. These disturbances can propagate by radiation and induction (through free space) and by conduction (over wires, chassis, and the like). Accordingly, one has radiation and induction noise in the former case, and conduction noise in the latter.

Radio noise may further be classed into in-system (or internal) and out-of-system (or external). Both types may be natural and man-made. As their name implies, out-of-system disturbances are produced by sources external to a given communication system. These disturbances include atmospheric noise, man-made noise, galactic (cosmic) noise, thermal radiation from the Earth, and interference produced by other radio stations intentionally or unintentionally. In-system disturbances may arise during the operation of a group of radio links which are all part of the same communication system and in the functional elements of a radio link itself. These disturbances include component noise, quantization noise, cross-coupling between channels, out-of-band and spurious radiation from transmitters, local-oscillator emission, switching and contact noise, a.c. hum, and some others.

Most forms of noise are nonstationary, random processes. There may, however, be deterministic disturbances, for example, harmonics, and stationary disturbances, such as thermal noise originating in the equipment and galactic (cosmic) noise. Depending on the extent of the wanted and noise signals along the time and frequency axes, the various forms of noise may be classed into three broad groups: time-concentrated (impulse), frequency-concentrated, and fluctuation noise.

Impulse noise is a nonstationary, random process due to disturbances having abrupt changes and of short duration. It consists of a nonperiodic, time-random sequence of discrete pulses whose duration, $T_{\text{i.n.}}$, is shorter than that of an elementary signal unit, T_s, and whose spectrum width, $F_{\text{i.n.}}$, is greater than that of the wanted signal, F_s. This type of noise is produced by lightning discharges, industrial equipment, and similar sources. It is random in terms of pulse height (or amplitude) $V_{\text{i.n.}}$, pulse duration $T_{\text{i.n.}}$, quiet intervals $\Delta T_{\text{i.n.}}$ between pulses, and their grouping into trains (each train will usually have a random number of pulses and the intervals between the trains are likewise random). The concept of impulse noise is associated not only with the parameters (amplitude and duration) of the noise, but also with those of the transients it causes in the selective stages of a receiver. Among other things, the noise appearing at the receiver output will remain to be of the impulse type if $T_{\text{i.n.}}$ is shorter than the transient time, T_t, and the interval $\Delta T_{\text{i.n.}}$ is longer than $T_{\text{i.n.}}$. Otherwise, it will turn into fluctuation noise at the receiver output.

Lightning discharges occur at many localities of the world at the same time. For this reason, a receiver may pick up noise from both nearby and distant thunderstorms. Characteristically, the spectral density of the noise from nearby thunderstorms varies in inverse proportion to frequency, whereas the noise from distant thunderstorms sets up a quasi-stationary, fluctuating field whose intensity varies with time of day, season of the year, and terrain, etc. In daytime, the noise level is 15 to 20 dB lower than it is at night; over the ocean it is 5 to 10 db lower than over land; finally, it is higher in summer than in winter. This form of disturbance has a maximum intensity at frequencies up to 30 kHz and decreases with increasing frequency at a rate of 50 dB per decade.

Frequency-concentrated noise is that whose frequency spectrum does not exceed the signal spectrum in width. This is, for example, true of unintentional interference from other transmitters, which is often the cause of poor reception.

Fluctuation noise is not concentrated in either time or frequency; it has a wider frequency spectrum than the signal, and it is always present at the receiver input and output. This form of noise includes internal equipment noise, galactic noise, and the overall contribution from a great number of concentrated or impulse components equal in intensity and having identical stochastic properties.

Galactic (cosmic) noise combines background noise and radio static the origin of which is due to sources outside the Earth's atmosphere (the planets, radio stars, and radio nebulae). The intensity of radiation from discrete sources depends on the directivity of the receiving antenna and the elevation above the radio horizon. With highly directional antennas, the background noise temperature is practically independent of the coordinates of the source; it decreases with increasing frequency at a rate of about 20 dB per decade, so that at frequencies in excess of 3 to 5 GHz it becomes negligibly small. At those frequencies, however, an increasing contribution comes from atmospheric noise due to the absorption and re-radiation of radio-wave energy by oxygen and water vapour; its frequency dependence has peaks at wavelengths of 0.25, 0.5 and 1.35 cm. In the frequency range from 1 to 10 GHz and at an antenna elevation of over 5° the galactic noise temperature, $T_{n,g}$, is anywhere between 1 and 3K, the atmospheric noise temperature, $T_{n,a}$, is 2 to 200 K, and the solar noise temperature, $T_{n.s}$, ranges from 2×10^4 to 3×10^5 K (and follows an 11-year cycle of variations). The terrestrial noise temperature, $T_{n,e}$, is negligibly low because thermal radiation from the Earth is solely intercepted by the side and rear lobes of the directional pattern of an antenna. That is why this frequency band is especially good for space communications.

The instantaneous value of fluctuation noise is normally distributed and has an expectation (mean value) of zero, whereas its phase is uniformly distributed. The probability density of the fluctuation noise envelope obeys the Rayleigh distribution

$$W(V_f) = (V_f/\sigma_f^2) \exp(- V_f^2/2\,\sigma_f^2)$$

The noise-envelope variance which characterizes the fluctuation intensity in the post-detector section of a receiver is

$$\mathrm{Var}_f = \int_0^\infty V_f^2 W(V_f)\, dV_f - \left[\int_0^\infty V_f w(V_f)\, dV_f \right]^2 \approx 0.43\sigma_f^2$$

Also, fluctuation noise is a ‘smooth’ one as compared with impulse noise; its peak-factor is

$$K_{pf} = V_{f,\,\max}/\sigma_f = 3$$

where $V_{f,\,\max}$ is the maximum instantaneous value of fluctuation noise. To demonstrate,

$$P(V_{f,\,\max} > 3\sigma_f) = \int_{3\sigma_f}^\infty W(V_f)\, dV_f \approx 0.0013$$

The most characteristic form of fluctuation noise is *white noise*. This is a stationary, random process whose energy per unit bandwidth is constant and independent of the central frequency of the band. The name is taken from the analogous definition of white light. In the limit of zero width and infinite height (amplitude), the impulses making up white noise become delta-functions, and the autocovariance of the noise, is also a delta function,

$$\varnothing(\tau) = (N/2)\,\delta(\tau)$$

The quantity N appearing in the above equation is called the unilateral single-sided power spectrum or spectral density of the noise (in watts per hertz). Any two time sections of white noise, spaced any arbitrarily small distance apart, are uncorrelated. Realizations of such a disturbance are unpredictable, which fact makes it difficult to combat it. A white-noise model of fluctuation noise cannot be realized because the source would have to have an infinitely high power. At the output of a band-limited filter, however, the noise power is always finite, therefore such an idealization does not complicate calculations, and fluctuation noise may be approximated by white noise if its spectrum is uniform within the bandwidth of the receiver's input circuit. In contrast to impulse noise, when white noise exists at the receiver input, the rms noise voltage at its output is proportional to the square root of the bandwidth. The point is that the components of white noise are uncorrelated and add together in power, whereas the components of impulse noise are combined in phase, that is, in voltage.

7.2. Man-Made Noise

The contribution that man-made noise makes to overall radio noise has been growing in scope with progress in technology. Except for remote rural localities, its level exceeds that of natural disturbances. Man-made noise may have a discrete and a continuous spectrum. Sources of man-made noise having a discretespectrum include industrial, scientific and medical units, the local oscillators of receivers, the sweep (time-base) generators of cathode-ray oscilloscopes, and the like. Man-made noise having a continuous spectrum is produced by car ignition systems, electrical home appliances, fluorescent lighting, electrical power transmission lines and electronic systems that radiate spurious signals. Man-made noise can be carried by metalwork, such as pipelines, elevator cables, etc. This type of noise can have a distance range of units to thousands of meters. The strength of radiation noise is stated in terms of the field strength at a specified distance from the source (ordinarily, in the range from 1 to 300 m) and expressed in decibels referred to (that is, above or below) 1 μV/m (or dBμV/m for short). Noise strength is stated in terms of the current and voltage existing at the output terminals of the source. Noise is attenuated on propagating away from its source, and the attenuation increases with increasing frequency, so that at frequencies in excess of 30 MHz it may quite often be neglected.

Noise originating in industrial, scientific and medical equipment can propagate over a distance range of up to 30 km and affect adversely radio and TV broadcast, radio communications, etc.

The ignition systems of internal-combustion engines generate trains of strong pulses with a duration of 0.2 to 0.5 μs. Their spectrum extends out to 2 GHz and has peaks in the range of 30 to 150 MHz.

Noise due to electrical power transmission lines occupy the frequency band from 3 kHz to 300 MHz, being at its strongest at frequencies up to 30 MHz. The noise level rises rapidly as the line voltage increases over 40 to 70 kV, and the spectrum peak is then shifted towards the high-frequency end.

Home appliances, such as refrigerators, floor-polishing machines, and power tools, make up an especially troublesome group of noise sources for nearby receivers operating off the same supply line.

Special measures are usually prescribed to combat man-made noise. For each type of source, there is a maximum allowable level established by an applicable code of practice, and its enforcement is supervised by duly appointed inspectors. In the USSR, the maximum allowable level of noise for industrial, scientific and medical equipment is 105 to 115 dBpW (where dBpW stands for 'decibels referred to 1 picowatt') within the allocated frequency range and 50 to 57 dBpW outside that range. For electric motors operated indoors in residences the figure is set at 45 to 55 dBpW in the frequency range between 0.15 and 300 MHz, and 55 dBpW in the frequency range from 300 to 1000 MHz.

7.3. Unintentional Radio-Station Interference

Interference from nearby transmitters is a major form of disturbance. In a receiver, it may give rise to undesirable nonlinear effects and go through directly over the adjacent and image channels.

Electromagnetic radiation from transmitters may be classed into wanted and unwanted. The wanted radiation contains the spectral components lying within the least bandwidth, B_{req}, required for the transmission of messages with the fidelity and rate specified for a given emission (or transmission) type. The unwanted radiation is made up of the components lying outside B_{req}. It may reach a receiver via its antenna, over the supply and switching circuits, through access holes in the chassis and electromagnetic shields, thus impairing its noise immunity. The unwanted radiation may further be classed into the out-of-band group and the spurious group.

The out-of-band group occupies bands adjacent to the band of frequencies associated with the wanted radiation, and its components owe their origin to the modulation of the wanted signal. This form of interference is combatted not only in the receiver but also in the transmitter by better signal filtering, by limiting the bandwidth of the modulating voltage, and by maintaining the linearity of the modulation characteristic.

The spurious radiation is due to nonlinear processes which take place in the transmitter elements and are not associated with modulation, and it can arise at frequencies markedly differing from the carrier, f_s. This is, for example, true of radiation at frequencies which are harmonics and subharmonics of the carrier.

The unwanted radiation of transmitters impairs the electromagnetic compatibility of radio facilities and is therefore subject to regulations.

Radiation from the local oscillators of receivers should also be considered, as it may find its way into free space through the antenna of a receiver, hook-up wires, and chassis. Taking an applicable Soviet standard as an example, the power of radiation from the local oscillators of HF receivers should not exceed 1.5 nW.

7.4. Noise Immunity of Radio Reception as a Complex Problem

Noise immunity is a major problem in the theory and practice of radio reception. It includes an analysis of receiver susceptibility to radio noise and interference, the search for signal reception techniques that could provide for the best noise immunity with a given class of signals and noise, and the selection of components and devices to implement such techniques.

Message reception in the presence of noise is a typically statistical problem. The decision device determining what message the received signal carries is a major functional element of any receiver. The optimization of this circuit is the subject-matter of optimal reception theory, which is a division of statistical communication theory. The primary objectives of optimal reception theory are to formulate an applicable decision rule, to synthesize the decision device subject to the chosen optimality criterion and specified initial constraints, and finally to analyse the performance of the decision device in quantitative terms.

The multitude of signal types used in telecommunication systems inevitably affects the choice of an approach to the task of optimal reception. A set of discrete signals is countable and finite, whereas a set of continuous signals is uncountable and infinite. Quite obviously in the latter case one should differentiate between the transmission of magnitude and the transmission of waveform. The former case has to do with the discrete transmission of continuous messages, where the magnitude of the information-bearing parameter remains unchanged for the duration of the signal, but it is a random variable. The latter case is the analog transmission of telephone messages, where the information-bearing parameter is a random function of time.

In communication theory, the reception of signals in the order given above is referred to as *signal extraction* (if one is concerned with the presence or absence of one or more signals, one speaks of *signal detection*), signal parameter estimation, and message reproduction. There is no fundamental difference between the above techniques of signal reception. Still, there are some differences in the approach and mathematical tools used when one has to handle an optimization problem. In more detail, this matter is discussed in a course on the theory of signal transmission.

For decision-making, one needs a definite minimum of prior data about signals and noise in the form of one- or multi-dimensional probability distributions, $p(x)$, or probability densities, $W(x)$. The signal parameters varying in a manner decided by the message being transmitted are information-bearing parameters,' their magnitudes are unknown at the receiving end, and their distribution is determined by the *apriori* distribution of messages. If in the received signal only information-bearing parameters are unknown, the signal is said to be known exactly. If, on the other hand, all the remaining parameters are likewise unknown, one has a signal with unknown parameters. The latter may be random variables or random functions of time; one speaks then of a signal with random parameters.

In the case of signals known exactly, a complete knowledge of *apriori* data about messages ensures the highest validity of the decision and a maximum noise immunity in reception. This is known as the *ultimate noise immunity*; it is realized by what is referred to as an ideal receiver. In practice, there is always some uncertainty about signals and noise, and the real noise immunity is lower than the ultimate noise immunity.

The theory of statistical decision-making used in tackling the problem of optimal reception is a theory of decision-making under uncertainty. Its fundamental concepts are the *loss* or *cost function* and the *average risk*. The task is to form the decision

$$\gamma = f[z(t)]$$

as to which signal has been transmitted by analysing the mixture $z(t)$ of a signal $x(t)$ and noise $n(t)$. The form of the function $f[z]$ has a direct bearing on the decision rule to be used. The decision γ may differ from the true value of x owing to the action of noise. Therefore, any decision is accompanied by a risk which is a function of both x and γ. The function $C_n(x, \gamma)$, called the loss or cost function, depends on the purpose to be served by a given radio link and on the requirements to be met by the received messages. The rule for its choice consists in that it should fit the problem at hand in the best possible way and increase with increasing difference between the estimate and true value of the signal.

Since x and γ are random arguments, the loss function C_n is likewise random. Therefore, in order that it can be used to describe the quality of reception, it is usual to introduce the mathematical expectation, or mean value, of the loss function

$$R_0 = E\,[C_n\,(x,\ \gamma)] \tag{7.1}$$

referred to as the average risk. An optimal receiver is that which minimizes R_0 with the specified form of the loss function C_n. From the view-point of statistical decision theory, receiver optimization reduces to finding a decision rule which would minimize the average risk. All such rules are known as *Bayes decision rules.*

For continuous signals the average risk is

$$R_0 = \int\limits_{S_x}\int\limits_{S_\gamma} C_n\,(x,\ \gamma)\,W\,(x,\ \gamma)\,dx\,d\gamma \tag{7.2}$$

where S_x and S_γ are the regions of likely values of x and γ.

For discrete signals,

$$R_0 = \sum_{i=1}^{N}\sum_{j=1}^{M} C_n\,(x_i,\ \gamma_j)\,p\,(x_i,\ \gamma_j) \tag{7.3}$$

where N is the number of likely signals and M is the number of likely decisions. In our further discussion, we will assume that $M = N$.

Example 7.1. Find R_0 for a symmetrical binary channel on the assumption of equiprobable messages. On the basis of Eq. (7.3), we have

$$\begin{aligned} R_0 &= p\,(x_1,\ \gamma_1)\,C_n\,(x_1,\ \gamma_1) + p\,(x_2,\ \gamma_1)\,C_n\,(x_2,\ \gamma_1) \\ &\quad + p\,(x_1,\ \gamma_2)\,C_n\,(x_1,\ \gamma_2) + p\,(x_2,\ \gamma_2)\,C_n\,(x_2,\ \gamma_2) \end{aligned}$$

Since

$$\begin{aligned} &p\,(x_i,\ \gamma_j) = p\,(x_i)\,p\,(\gamma_j \mid x_i) \\ &p\,(x_1) = p\,(x_2) = 0.5 \\ &p\,(\gamma_1 \mid x_1) = p\,(\gamma_1 \mid x_1) = q_0 \\ &p\,(\gamma_1 \mid x_2) = p\,(\gamma_2 \mid x_1) = p_0 \end{aligned}$$

where q_0 and p_0 are the probabilities of correct and false reception, it follows that

$$\begin{aligned} R_0 &= 0.5\,q_0\,[C_n\,(x_1,\ \gamma_1) + C_n\,(x_2,\ \gamma_2)] \\ &\quad + 0.5\,p_0\,[C_n\,(x_1,\ \gamma_2) + C_n\,(x_2,\ \gamma_1)] \end{aligned}$$

Depending on the applicable decision rule, we obtain p_0 and q_0 and, for the specified form of C_n, the value of R_0.

Example 7.2. In signal parameter estimation and in continuous message reproduction, it is often practised to use the quadratic loss function

$$C_n\,(x,\ \gamma) = (\gamma - x)^2 \tag{7.4}$$

On substituting (7.4) in (7.2), we get

$$R_0 = E[(\gamma - x)^2]$$

that is, the average risk is the mean-square error. Therefore, the often used criterion in the form of a minimum mean-square error is a special case of the minimum R_0 criterion when the loss function has the form of (7.4).

Example 7.3. In the transmission of discrete signals, it is customary in radio communications to use the simple loss function

$$C_n(x_i, \gamma_j) = \begin{cases} C_0 = 1 & \text{for } i \neq j \\ 0 & \text{for } i = j \end{cases} \qquad (7.5)$$

Then, in agreement with (7.3), we have

$$R_0 = C_0 \sum_{i=1}^{N} p(x_i) \sum_{j=1}^{M} p(\gamma_j | x_i) = C_0 P_{\text{error}} \qquad (7.6)$$

where P_{error} is the average probability of error in reception. Thus, with the simple loss function, an optimum receiver minimizes the error probability. Such a receiver is called the Kotelnikov-optimal or ideal-observer receiver, and the optimality criterion defined by Eq. (7.6) is known as the Kotelnikov criterion or the Ideal-Observer criterion.

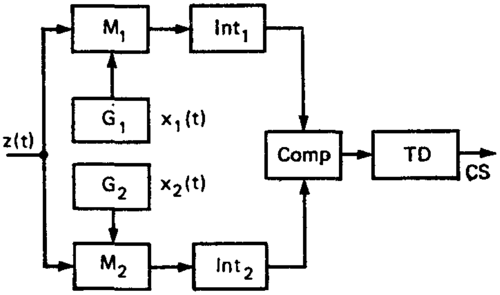


Fig. 7.1

As we choose different loss (or cost) functions, we obtain different optimality criteria all of which are the special cases of a common criterion—the minimum average risk.

The arrangement of an ideal binary-signal receiver which has the ultimate white-noise immunity is shown in Fig. 7.1. It is seen to contain expected-signal generators G_1 and G_2, multipliers M, integrators *Int*, a comparator *Comp*, and a threshold device *TD* which generates a control signal, *CS*.

The multipliers do the job of detectors as they extract the envelope of the input signal. To demonstrate, if

$$x_i\ (t) = V_i \sin \omega t$$

then what we have at the output of the respective multiplier will be

$$x_i\ (t)\ x_i\ (t) = 0.5\ V_i^2 - 0.5\ V_i^2 \cos 2\ \omega t$$

The h.f. components are eliminated by the associated integrator which operates as a low-pass filter. For its operation, a detector in the form of a multiplier needs a reference voltage synchronized in frequency and phase with the incoming signal. For this reason, it is called a *coherent detector*. Therefore, an ideal-observer receiver may alternatively be referred to as an averaging coherent receiver in which averaging is done by an integrator.

The above arrangement is often simplified, but this simplification entails an impairment in noise immunity.

The decision is made on comparing the voltages at the integrator outputs just as the signal ceases, that is, at $t = T_s$, which corresponds to the general concept of signal filtering. The integrators are often replaced by low-pass filters, and the signal is estimated at a time $t < T_s$, usually at the middle of the signal. This method is known as *gating* or the *single-sample method*. It is also known as the 'shortened-contact' method. Such a receiver is a non-averaging coherent receiver.

An important advantage of coherent reception is the fact that the output signal-to-noise ratio is a linear function of the input signal-to-noise ratio. With incoherent reception and weak signals, this relation is a quadratic one, and noise immunity is impaired.

The device can further be simplified if the coherent detector is replaced by an incoherent one. This yields a non-averaging incoherent receiver. It is the simplest of all, but its noise immunity is low. With a large input signal-to-noise ratio, a coherent detector offers, however, an insignificant advantage. Noise immunity may sometimes be improved by placing an integrator past the detector. This is reception with post-detector integration, and the receiver is an averaging incoherent one.

There may also be an incoherent receiver with pre-detector integration, in which case the integration is done in the i.f. section by a narrowband filter usually followed by a usual incoherent detector.

The multipliers and integrators in the arrangement of Fig. 7.1 make up between them what is known as a *cross-correlator*. In an optimum receiver, it not only performs signal detection and integration, but also signal separation. It is desirable that when only one signal, say, $z\ (t) = x_1\ (t) + n\ (t)$ exists at the input, the output voltage of the cross-correlator should be present in the arm assigned to x_1, and no voltage should exist in the other arm. To achieve this, one chooses the signals to be of such a form that their cross-correlation is a minimum. Then the cross-correlator will separate the signals.

Let a telecommunication system use orthogonal signals

$$x_1(t) = V_1 \sin \omega t$$

and

$$x_2(t) = V_2 \cos \omega t$$

Then, in the transmission of the signal $x_1(t)$, we will have at the output of multipliers M_1 and M_2 in the receiver of Fig. 7.1 the following:

$$x_1^2(t) = 0.5\, V_1^2 (1 - \cos 2\,\omega t)$$

$$x_1(t)\, x_2(t) = 0.5\, V_1 V_2 \sin 2\omega t$$

That is, there will be no low-frequency signal in the second arm, and the spurious detection products will have been discarded by the integrator.

The replacement of a coherent detector by an incoherent detector removes the ability to separate signals. Therefore in an incoherent receiver the detector must be preceded by separating devices. In a binary FSK receiver, these devices may be filters which separate 1s from 0s.

With all other conditions being equal, the noise immunity of an optimum receiver depends solely on the ratio of signal energy to the noise spectral density. It is independent of the frequency band occupied by the signals because the passband of an optimum receiver perfectly matches the width of the signal spectrum. As the passband is increased, the number of noise components increases, but so also does the number of signal components. In a real receiver, the error probability is determined by the ratio of signal power to noise power which is proportional to the passband.

The evaluation of noise immunity in signal reception is complicated when, instead of an isolated disturbance, one has to consider a complex of disturbances. If a receiver only contained linear elements, the problem would reduce to determining the response of the receiver to each component disturbance. However, real receivers contain nonlinear elements as well, and their response to the signal and noise gives rise to further components, and the statistical characteristics of the processes are changed.

Example 7.4. An additive mixture, $z(t)$, of a wanted signal $x(t)$, a concentrated interfering signal $y(t)$ and noise $n(t)$ is applied to the input of an FM receiver. Let $x(t)$ and $y(t)$ have an FM waveform:

$$x(t) = V_s \cos \alpha$$

and

$$y(t) = V_d \cos \beta$$

where

$$\alpha = \omega_s t + \Delta\omega_s \int_{-\infty}^{t} s_s(t)\, dt$$

$$\beta = \omega_i t + \Delta\omega_i \int_{-\infty}^{t} s_d(t)\, dt$$

where

$s_s(t)$ = modulation process of the wanted signal
$s_i(t)$ = modulation process of the interfering signal
$\Delta\omega_s$ = wanted-signal frequency deviation
$\Delta\omega_i$ = interfering-signal frequency deviation

This case in which the wanted and interfering signals are alike is most dangerous. Let us adopt the following assumptions for the receiver: the gain of each stage is unity, both the frequency detector and the output low-pass fillter are ideal, and the numerical characteristic of noise immunity is the signal-to-interference power ratio,

$$h_s^3 = P_s/P_i \quad .$$

Let us denote $\eta = V_i/V_s$, $\omega_{dif} = |\, \omega_i - \omega_s \,|$, and $\psi = \beta - \alpha$. Then the resultant waveform produced by $x(t)$ and $y(t)$ will be

$$v_0 = V_0 \cos(\alpha + \theta)$$

where

$$V_0 = V_s(1 + \eta^2 + 2\eta \cos\psi)^{1/2}$$

and

$$\theta = \text{arc tan}\, [\eta \sin \psi/(1 + \eta \omega s \psi)]$$

Assume also that $P_s \gg P_i$ and $P_s \gg P_n$. Then the response of the frequency detector to the noise and the interfering signal may be analysed separately, and the spectrum at the detector output can be found by adding together the spectra of the interfering signal and the noise. Recalling that $\eta \ll 1$ and considering the expressions for V_0 and θ, we get

$$\theta = \text{arc tan}\, (\eta \sin \psi)$$

and

$$v_0 = V_0 \cos[\omega_s t + \varphi(t)]$$

where

$$\varphi(t) = \Delta\omega_s \int_{-\infty}^{t} s_s(t)\, dt + \theta(t)$$

The output signal of an ideal frequency detector is proportional to the derivative of the phase of the input waveform, where the proportionality factor is the slope of the characteristic curve of the frequency detector, S_{FD}. Then

$$v_{\mathrm{out}}(t) = S_{\mathrm{FD}}\, d\varphi\,(t)\, dt = S_{\mathrm{FD}}\ [\Delta\,\omega_s s_s\,(t) + d\theta\,(t)/dt]$$

Here, the first term is due to the wanted signal and the second, to the interfering signal. That is,

$$v_{\mathrm{out}}(t) = V_{s,\mathrm{out}} + V_{\mathrm{i,out}}$$

The wanted-signal output power of the frequency detector is

$$P_s = S_{\mathrm{FD}}^2\,\Delta\,\omega_s$$

and the interfering-signal output power is

$$P_{\mathrm{i}} = (1/2\pi)\int_0^{2\pi B_{\mathrm{F}}} G_{1,\,\mathrm{out}}(\omega)\, d\omega$$

where the interfering-signal power spectral density $G_{\mathrm{d}}\,(\cdot)$ depends on the spectral densities $G_{s,m}\,(\omega)$ and $G_{\mathrm{i},m}\,(\omega)$ of the processes

$$\cos\left[m\Delta\omega_s \int_{-\infty}^{t} s_s(t)\, dt\right]$$

and

$$\cos\left[m\Delta\omega_{\mathrm{i}} \int_{-\infty}^{t} s_{\mathrm{i}}(t)\, dt\right]$$

The noise power at the detector output in the passband of the low-pass filter is

$$P_{\mathrm{n}} = (1/2\pi)\int_0^{2\pi B_{\mathrm{f}}} G_{\mathrm{n}}(\omega)\, d\omega$$

where $G_{\mathrm{n}}\,(\omega)$ is the noise power spectral density.

In order to determine the overall disturbance power at the detector output

$$P_{\mathrm{overall\text{-},out}} = P_{\mathrm{i}} + P_{\mathrm{n}}$$

it is necessary to find $G_{s,m}\,(\omega)$, $G_{1,\,m}\,(\omega)$, and $G_{\mathrm{n}}\,(\omega)$. In the general case, the expressions for these quantities reflect the effect produced by the nonlinear interaction of $x\,(t)$, $y\,(t)$, and $n\,(t)$ in the frequency detector.

Example 7.5. A phase detector is fed an additive mixture

$$z\,(t) = x\,(t) + y\,(t) + n\,(t)$$

where the wanted signal is

$$x\,(t) = V_s \cos\,[\,\omega_s t + \varphi_s\,(t)]$$

Within the $(0, 2\pi)$ interval, the signal epoch may take on fixed values

$$\varphi_{s,j} - \varphi_{s(j-1)} = 2\pi/2^k$$

where $k = 1, 2, \ldots$ For simplicity, let $k = 1$, which corresponds to 180-deg phase-shift keying. Assume that the detector is an ideal one, with zero width of the thresholds defining the limits of changes in the signal elements. Should their vector cross the boundary under the action of the interfering signal or noise the received signal element will be a false one. Since $k = 1$, then $\Delta\varphi = \pi/2$. The concentrated interfering signal is

$$y(t) = V_i \cos[\omega_i t + \varphi_i(t) + \varphi_0(t)]$$

where $\varphi_i(t)$ = phase modulation of the interfering signal
$\varphi_0(t)$ = epoch uniformly distributed over the $(0, 2\pi)$ interval

The receiver noise may be written as

$$n(t) = V(t) \cos \omega_s t + U(t) \sin \omega_s t$$

In coherent reception, one input of the phase detector accepts a mixture, $z(t)$, and the other, the reference voltage from the local oscillator

$$v_{LO}(t) = V_{LO} \cos \omega_s t$$

The voltage appearing at the output of the low-pass filter is

$$V_0 = V_s (X_0^2 + Y_0^2)^{1/2}$$

where

$$X_0 = V_s + V_i \cos[\omega_{\text{dif}} t + \gamma(t)]$$

and

$$Y_0 = V_i \sin[\omega_{\text{dif}} t + \gamma(t)] + u(t)$$

Here,

$$\gamma(t) = \varphi_i(t) + \varphi_0(t)$$

An error will arise if

$$|\alpha_0| \geqslant \Delta\varphi$$

where α_0 is the phase of the resultant vector.

It is convenient to construct the noise immunity characteristic as a plot of $P_{\text{error}}(h_s^2, k^2)$, where h_s^2 is the signal-to-noise ratio, and k^2 is the allowable interference-to-signal power ratio at the receiver input. This plot appears in Fig. 7.2.

When one analyses a receiver for noise immunity, one usually assumes that the disturbance finds its way into the receiver via its antenna. Actually, it is not always the case. A communication radio receiver is ordinarily built as a collection of functional subunits which have a common antenna input, one or several power supply units, and a common grounding bus. In the circumstances, disturbances may reach the various subunits over the supply circuits, access holes in the shields, grounding circuits, and so on, rather than via the antenna. Therefore, the action of disturbances may be depicted as shown in Fig. 7.3 where $y_i(t)$ designates the additive or multiplicative effect of the disturbance on the signal at the input to the ith subunit described by an operator L_i. Analysis of the receiver for noise immunity in such conditions enables one to pin-point the most vulnerable spot in the system, to determine the required noise protection ratio at each point in the receiver, to find the admissible set of noise signals, to specify their limits, etc. Among the factors that affect noise immunity are the properties of the signals and noise, their significance in relation to the performance of the receiver subunits, the manner of signal processing, the presence of nonlinear elements, the likely paths of noise ingress, and so on. Since the problem is an extremely complex one, it is usual to limit oneself to an approximate estimation of noise immunity. After one has chosen models to represent the wanted and noise signals involved, the estimation of noise immunity reduces to determining the order of magnitude and the trend of changes in the numerical characteristic of noise immunity with the signal and noise parameters. It is simpler to form an estimate of noise immunity than to calculate it exactly, as coarser mathematical models will then usually do.

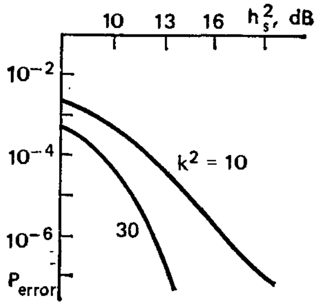


Fig. 7.2

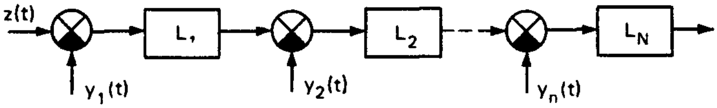


Fig. 7.3

7.5. A General Outline of Noise and Interference Control in Radio Receivers

Noise and interference in a radio receiver can be controlled in more than one way. Consider some of thę techniques.

When transmitters are located close to receivers, as is the case on board ships, aircraft, and at radio centres, the emf induced in a receiving antenna may be as high as 100 volts or even more. Because of this, there is a danger of damage to be done to the first stage in the r.f. amplifier. As an illustration of how this danger can be avoided, Fig. 7.4 shows a simplified schematic of the preselector used in a point-to-point HF infradyne receiver (see Sec. 4.5). The allowable emf at the input it 100 volts; should it exceed the limit, a relay, *Rel*, will operate and connect the antenna input to ground. There is an anti-radar low-pass filter, *ARF*, which has a cutoff frequency of

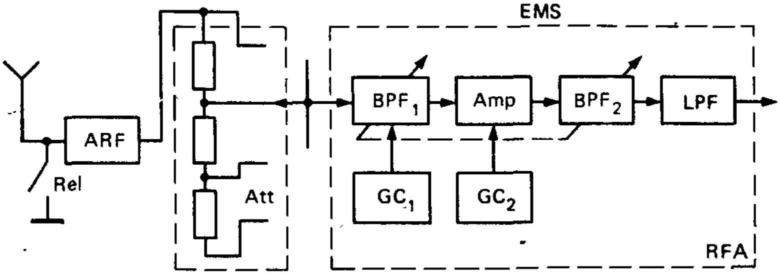


Fig. 7.4

around 200 MHz. As its name implies, it is intended to protect the receiver input against strong signals from radars operating at higher frequencies. In the attenuator, *Att*, the input voltage is attenuated in steps of 10, 20, or 30 dB.

The attenuator is brought in circuit in the reception of strong signals. As it attenuates the wanted signal to an acceptable level (just above the sensitivity threshold), it proportionately attenuates all the interfering signals whose number and total level are high in the wide passband of the preselector. The attenuation of interfering signals reduces the intensity of spurious responses produced by the interaction between themselves and with the incoming signal in the early electron devices of the receiver.

This interaction has two objectionable consequences:

— the stronger signals cause changes in the weaker signals. Notably, the incoming signal modulated by the intelligence being transmitted may at the same time be modulated by various disturbances;

—the nonlinear conversion of a complex overall spectrum produces intermodulation components which are close in frequency to the incoming signal and fall within the bandwidth of the receiver. Being superimposed on the signal, they can irrecoverably corrupt it.

The spurious responses produced by nonlinear conversion are proportional to the 2nd, 3rd and higher powers of the voltages associated with interfering signals. Therefore, their reduction mitigates their effect manyfold. If, for example, the attenuator brings down the signal voltage at the input to a receiver by a factor of 10 to 20, the spurious responses produced by the above events will be suppressed by a factor of several hundred.

The receiver contains a bandpass filter, BPF_1, which includes a guard circuit, GC_1, with an operating threshold of about 20 volts. The amplifier, *Amp*, that follows the bandpass filter is built around a FET with a large dynamic range and also includes a guard circuit of its own, GC_2. The amplifier load is a second bandpass filter, BPF_2, with a cutoff frequency of 31 MHz. This filter provides image rejection, with the first i.f. exceeding 31 MHz.

The electromagnetic shield, *EMS*, serves to attenuate the electromagnetic fields set up by extraneous sources. Unfortunately, shielding cannot keep out all of the man-made interference if a source of interference is connected to the receiver by wires. In such a case, the interference-carrying wires should include rejection filters in order to block the passage of r.f. currents and to pass direct current and currents at power frequencies without noticeable attenuation.

Other techniques of interference and noise control are related to signal processing. Among them is a special group based on signal discrimination or selection.

A signal has a finite set of parameters; the qualitative and quantitative differences between them can be used for purposes of signal discrimination or selection. Thus, there may be space discrimination, polarization discrimination, frequency discrimination, time discrimination, amplitude selection, statistical selection, and functional selection.

Signal discrimination or selection may be further characterized in terms of the volume of information used: it may be based on the processing of individual signals or a group of signals arriving over one or several channels. There may also be primary discrimination or selection and secondary discrimination or selection, according to the nature of prior knowledge available. Primary selection is based on differences between the descriptive parameters of signals and interference. Secondary selection uses special features added to a radio signal in order to enhance the accuracy of selection.

7.6. Disturbance Suppression by Cancellation

Signal and disturbance spectra often overlap — a fact which stands in the way of frequency discrimination of signals so that one has to use other discrimination or selection techniques.

Channels can be separated in space and disturbance signals can cancel out, if one uses two receiving antennas, as shown in Fig. 7.5. To demonstrate, in the case of a concentrated disturbance signal arriving at the two antennas, A_1-and A_2, we have

$$v_{d1}(t) = V_{d1} \exp [j (\omega t + \varphi_0 + \varphi_d)]$$

and

$$v_{d2}(t) = V_{d2} \exp [j (\omega t + \varphi_0)]$$

where φ_0 is the epoch and

$$\varphi_d = (2\pi d/\lambda) \sin \theta_d$$

is the phase difference determined by the spacing d between the antennas and the angle of disturbance arrival at the antennas, θ_d. Then, at the output of the subtractor, *Sub*, we have

$$x_d (t) = v_{d1} (t) K_1 \exp (- j\varphi_1) - v_{d2} (t) K_2 \exp (- j\varphi_2)$$

where φ_1 = phase shift produced by phase-shifter PS_1
φ_2 = phase shift produced by phase-shifter PS_2
K_1 = gain of Amp_1
K_2 = gain on Amp_2

If the conditions $| \varphi_1 - \varphi_2 | = \pi$ and $K_1V_{d1} = K_2V_{d2}$ are satisfied, no disturbance will appear at the subtractor output. If arrival of a disturbance signal is accompanied by arrival of a wanted signal, but from a different direction such that $\theta_s \neq \theta_d$, then $\varphi_s \neq \varphi_d$, and there will be a wanted signal voltage at the subtractor output. Unfortunately, the need to have two antennas and the requirement for exact tracking of two phase shifters (so that $\varphi_1 - \varphi_2 = \pi$) make this device rather difficult to build. Still worse, it becomes inefficient in the presence of several interfering signals arriving from different directions.

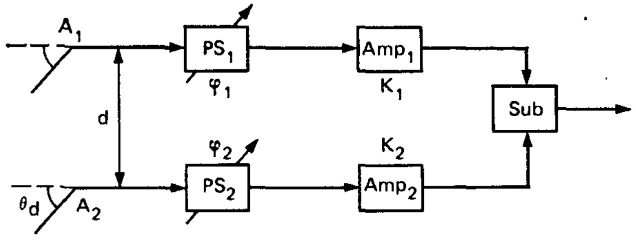


Fig. 7.5

If an interference is produced by a transmitter located close to the receiver, it can be cancelled out by a device such as shown in the block diagram of Fig. 7.6. Here, *Xmtr* is the offending transmitter. The reference voltage, V_{ref}, equal in magnitude to the interfering voltage, is fed by an auxiliary antenna to a bucking voltage generator, *CVG*. This unit contains two linear amplifiers, Amp_1 and Amp_2, with gains K_1 and K_2, respectively. The reference voltage is then inversed in phase so that it is in anti-phase with respect to the interference that reaches the input of the receiver, *Rvr*, from its own antenna, A. To effect this phase inversion, the circuit of Amp_1 contains a 90-deg phase shifter, PS. The voltage V_1 from the phase shifter is combined with the voltage V_2 from Amp_2 in a summator,

Σ, to produce the bucking voltage. The envelope detector, *ED*, detects the instant when thė voltage at the output of the subtractor, *Sub*, crosses zero, and generates a control signal proportional to the uncancelled (residual) interference. This signal goes to a control device, *CD*, which causes K_2 to change in proportion.

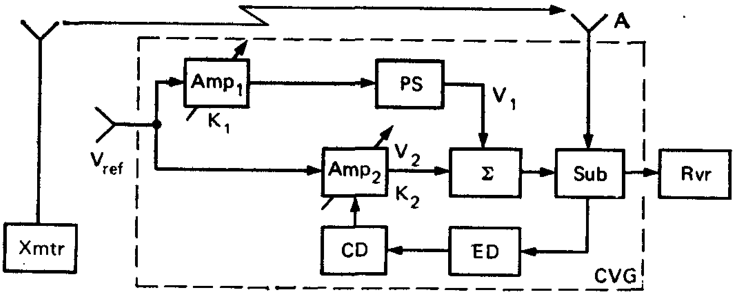


Fig. 7.6

Special promise is held by adaptive devices which include a correlator to process the incoming waves, and a microprocessor to control the amplitude, frequency and phase of the bucking voltage fed to the receiver input. Such devices can handle interfering signals from several sources at a time.

7.7. Receiver Protection Against Overload due to Disturbances

The upper limit for the input-signal strength that can be handled by a receiver is set by the nonlinear volt-ampere characteristic of the amplifying and frequency-converter (mixer) stages, whereas the lower limit is set by the receiver noise level. The two limits define the dynamic range of the receiver (see Sec. 1.10). Since the signal level may change by as much as 100 to 200 dB, it is of primary importance to extend the dynamic range. With the bandwidth chosen for a given class of signals, the dynamic range can only be extended upwards. This is achieved through the use of electron devices having low nonlinearity, low-noise BJTs and FETs, balanced frequency converters based on Schottky barrier diodes, negative feedback, selective stages with a gain reduced to a value ensuring the specified sensitivity, and adjustable attenuators at the receiver input (see Sec. 7.5).

The advantages offered by the infradyne (see Sec. 4.5) have led to the wide use of this receiver typẹ for andio communication in the HF band. The infradyne principle is aulso utilized in broadcast receivers with digital injection frequency synthesizers. Since in such receivers the preselector has a wide passband (tens of megahertz), it is inevitable that the signal should be accompanied by a wide spectrum of high-level interference and noise, a fact which calls for the input stages to be highly linear. Apart from the development of novel highly linear broadband input amplifiers and frequency converters, one has to look for other ways and means of interference and noise control. This basically consists in reducing the passband of the input circuits by means of switchable filters or in the suppression of the strongest concentrated disturbances by rejection filters servoed by an automatic search control system. Automated devices of this kind are now quite feasible on the basis of microprocessors.

7.8. Space and Polarization Signal Discrimination

Space signal discrimination can be based on antennas which have a suitable radiation pattern and are properly oriented, or on the functional (linear or nonlinear) processing of signals coming from several antennas. Various radiation patterns can be formed and the antennas can be oriented as may be required with special ease in the case of phased arrays. Space signal discrimination is best of all effected in the SHF band. On HF radio links less than 3000 km in length, the difference in the angle of beam arrival is anywhere between 15 and 20 deg, whereas for the longer links the figure is 5 or 6 deg. This complicates the use of highly directional antennas with a radiation pattern controlled in a vertical plane.

Polarization signal discrimination is based on the difference in polarization between the wanted signal and interference fields. The job of a polarization discriminator is done by the receiving antenna-feeder system, and the output power depends on how closely its polarization characteristic comes to that of the incoming waves. Polarization discrimination improves the detection of wanted signals in the midst of unintentional station interference and reflections from atmospheric discontinuities. The use of rotational polarization in HF and lower-frequency radio links is limited by the difficulties in building suitable antennas. Also, these frequency bands show a relatively strong depolarization of radio waves by the terrestrial magnetic field.

7.9. Frequency Signal Discrimination

Frequency discrimination refers not only to the simple separation of the signal and interference having non-overlapping spectra by suitable filters, but also to the extraction of the wanted signal from its mixture with an interference on the basis of the differences in their frequency spectra.

The suppression of strong concentrated disturbances calls for a preselector with an improved selectivity, that is, one having a greater number of resonant circuits and a high *Q*. When cooled by liquid helium, resonant circuits may have a Q of the order of 3.5×10^5 to 6×10^5 in the HF band, so that the interference can be attenuated by as much as 50 to 70 db within a percentage bandwidth of 1%. Unfortunately, cryogenic preselectors are complex in construction and not at all easy to operate and serve.

Two cases are of interest, namely overloaded and nonoverloaded frequency channels. Since interference sources are statistically independent, one may speak of the average channel loading density, $\alpha_{\mathrm{ch.l}}$, defined as the number of disturbances per unit bandwidth. This quantity offers a measure with which to evaluate the probable number of concentrated disturbances, N_{d}, within the bandwidth, B.

In overloaded channels, disturbances are very high in number. In the HF band, the number of disturbances with a level of 40 to 50 dBμV and more, registered at the point of observation, may be as great as 1000. This number includes tens of disturbances with a level of 90 to 100 dBμV. The loading density is anything but uniform: close on 93% of the total number is concentrated in the frequency range from 6 to 18 MHz.

Signal discrimination in the midst of disturbances which have overlapping frequency spectra can be effected by optimum matched filters. Quite often, the optimality criterion in such cases is the signal-to-interference ratio at the filter output. As the theory of signal transmission tells us, for this ratio to be a maximum the filter should have a transfer function of the form

$$K(j\omega) = cG^*(j\omega)\exp(-j\omega t_0)$$

where c is a constant and $G^*(\cdot)$ is the complex conjugate spectrum of the input signal. This two-port is 'optimum' in the sense that it maximizes the output voltage, and it is 'matched' in the sense that its transfer function matches the signal spectrum. To state this differently, the frequency response of the filter should coincide with the amplitude vs frquency characteristic of the signal accurate to within the constant, whereas the phase characteristic should be opposite in sign to the phase characteristic of the signal accurate to within the term ωt_0. With a transfer function like that, two events take place. Firstly, at the sampling instant the amplitudes of the harmonic components of the output signal are added together arithmetically so that the output voltage is a maximum. To demonstrate, the total phase of any spectral component of the signal at the filter output is

$$\varphi_0(\omega) = \omega t + \varphi_s(\omega) + \psi_0(\omega)$$

where $\varphi_s(\omega)$ = phase of a signal component at the input
$\psi_0(\omega)$ = phase response of the filter.

Since, however,

$$\psi_0(\omega) = -\omega t_0 - \varphi_s(\omega)$$

it follows that

$$\varphi_0(\omega) \equiv 0$$

for all $t = t_0$, irrespective of frequency, which means that at those instants the harmonic components of the signal combine. Secondly, the filter multiplies each harmonic component of the signal by the factor $K(\omega)$ which increases in magnitude with the amplitude ot the respective harmonic component. Since the noise intensity is the same at all frequencies, the best possible signal-to-noise ratio is obtained.

It follows from the two facts stated above that the signal appearing at the output of an optimum filter has a distorted waveform. That is why such filters can only be used for discrete messages in which case a receiver makes the decision as to the presence of a signal on the basis of the maxima of samples, whereas the waveform of the output signal is of no consequence. Such filters cannot be used for the reception of continuous signals; this job is done by filters which minimize the mean-square error in reproduction.

If a noise signal, $n(t)$, is present at the input to an optimum filter matched to a wanted signal $x(t)$, then, as follows from Duhamel's theorem, the output signal at time t_1 will be given by

$$v(t_1) = \int_0^{t_1} n(t_1)\, h_{\text{imp}}(t_1 - t)\, dt = c \int_0^{t_1} n(t)\, x(t_0 - t_1 + t)\, dt$$

where $h_{\text{imp}}(\cdot)$ is the impulse response of the filter. At $t_1 = t_0 = T_s$,

$$v(t_0) = c \int_0^{T_s} n(t)\, x(t)\, dt$$

This expression describes a short-term cross-correlation function. Thus, an optimally matched filter corresponds to the cross-correlator of an optimum receiver. Both types of receiver give the same noise immunity but contain different functional elements. A receiver containing an optimum filter effects coherent reception with integration, therefore, one needs to know the wanted signal accurate to within its epoch as before. Knowledge of the signal epoch is essential because the filter output is sampled at a precisely defined instant of time.

As a way of simplifying the device structure, the filter is usually followed by a conventional amplitude detector and the presence of a signal is established on the basis of its envelope. In consequence, the requirements for synchronization may be made more lenient because minor deviations from the time t_0 lead only to a slight reduction in the signal level. This is an incoherent receiver, and it is inferior in noise immunity to an optimum receiver.

7.10. Time Discrimination and Signal Averaging

The parameters of importance in time discrimination are the pulse duration (or width), the rise time, the fall (or decay) time, the pulse or pulse-group (code) repetition period, and the position of information-bearing pulses on the time axis relative to reference. Time discrimination mitigates the effect of impulse noise whose components have a shorter duration than the wanted signals, and also broadband asynchronous impulse noise.

Good results in the reduction of the threshold signal-to-noise ratio at a specified probability of correct reception are obtained when the time discrimination of pulse signals is combined with coherent averaging. Signal averaging is a special case of optimum filtering where the disturbance is white noise. In contrast to optimum matched filtering, signal averaging can be implemented by a summator or an integrator.

In order to illustrate the capabilities of signal averaging consider the case where the objêctive is to detect video pulses having a height (amplitude) V_s in the presence of noise $n(t)$. When use is made of a summator, a number N of samples is collected over the observation time T_{obs} from a mixture of the signal and noise

$$z(t) = x(t) + n(t)$$

If the intervals between adjacent pulses are greater than the noise correlation time, then the signal at the summator output will be

$$z_{out} = NV_s + \sum_{i=1}^{N} n_i$$

where n_i is the value of the function $n(t)$ at the instants when the amplitude of the ith pulse is sampled, t_i ($i = \overline{1, N}$). The second term (the total disturbance) is characterized by the variance $\sigma_d^2 = N\sigma_n^2$ where σ_n^2 is the variance of the function $n(t)$. Then the signal-to-noise ratio with signal averaging is

$$h_{s,\,av}^2 = NV_s^2/\sigma_n^2$$

Since without signal averaging the signal-to-noise ratio is

$$h_{no\text{-}av}^2 = V_s^2/\sigma_n^2$$

it follows that with signal averaging the signal-to-noise ratio is improved N times.

When signal averaging is done by an integrator, the output signal is

$$z_{out}(t) = V_s + (1/T_{obs}) \int_0^{T_{obs}} n(t)\, dt$$

The total noise component represented by the second term has a variance

$$\sigma_{\rm d}^2 \leqslant \tau\sigma_n/T_{\rm obs}$$

where τ is the correlation time of the noise. Therefore,

$$h_{\rm s,av}^2 \geqslant h_{\rm no\text{-}av}^2 T_{\rm obs}/\tau$$

Since the ratio $T_{\rm obs}/\tau$ is equal to the number N of uncorrelated samples of noise over the observation time $T_{\rm obs}$, it follows that both techniques are equivalent, but the one using an integrator is simpler to instrument.

7.11. Amplitude Selection

Three basic types of amplitude selector are used for noise suppression. The first type flattens negative peaks and is used to separate the wanted signals from impulse noise of a lower amplitude, thus improving the signal-to-noise ratio. The second types flattens positive peaks and is used to limit the amplitude of impulse noise. The third type is a combination of both and effects both positive and negative limiting.

Whereas white noise cannot be completely suppressed for fundamental reasons, impulse noise can be suppressed completely, at least theoretically. The practical techniques of impulse noise suppression may be classed into the cancellation type, the dynamic type, and the correction type. Impulse noise cancellation can be effected both ahead of and beyond the detector. Pre-detector cancellation uses a device which has two channels, one handling the total wave, and the other detuned from the signal carrier by a certain amount and supplying a cancellation or bucking voltage. By effecting frequency conversion and phase reversal, it is possible to generate at the output of the cancellation channel an impulse noise voltage which has the same parameters as the one in the total-wave channels and to subtract from the latter so as to cancel it. Unfortunately, the two channels of such a device must of necessity be almost perfectly identical in characteristics and, still worse, the cancellation channel itself may provide a path for impulse noise. Because of this, such devices have not found any appreciable use.

Correction techniques are based on the use of noise-immune codes with redundancy. Dynamic techniques use a combination of amplitude limiting and time discrimination.

Amplitude limiting is used in any one of several forms. In one of them, the i.f. amplifier section contains a broadband linear amplifier with a passband $B_{\rm b}$, an amplitude limiter, and a narrowband amplifier. That is why the technique is sometimes called the B-L-N method. The limiting level, V_{lim}, is chosen to be below the mean-square value of the overall signal plus noise voltage and is maintained constant.

Let the signal reaching a receiver be of duration T_s and let it also be accompanied by a short-duration disturbance of the impulse-noise type such that $T_n \ll T_s$. At the output of the linear section, the noise amplitude, V_n, is proportional to B_b, that is, an increase in B_b causes an increase in V_n as well. At the same time the impulse noise signal is made narrower, and the duration of the contaminated portion of the signal, t_{cont}, where $V_s < V_n$, decreases. The limiter equalizes the signal and the noise in amplitude. The narrowband amplifier for which the bandwidth-duration product, $B_n T_s$, is unity very nearly, suppresses the noise still more, whereas the signal has ample time to reach its steady-state value.

Let us evaluate the improvement in the signal-to-noise ratio, h_s^2, produced by the B-L-N structure. If the narrowband amplifier uses a single tuned circuit, the envelope of the output voltage will be

$$v(t) = V_0 [1 - \exp(-2B_n t)]$$

where V_0 is the steady-state value of the wave. Then for the instants at which the wanted and noise signals cease we have

$$v_s(T_s) \approx V_{s,\,\max}$$

and

$$v_n(T_n) \approx 2V_0 B_n T_n$$

whereas at the amplifier output

$$h_s^2 \approx B_b^2/4B_n^2 k_b^2$$

where

$$k_b = B_b T_n$$

Thus, the improvement in the signal-to-noise ratio provided by the B-L-N structure increases with an increase in the B_b/B_n ratio.

The presence of an amplitude limiter may have an adverse effect on noise immunity in the case of a concentrated disturbance. Let, for example, two disturbance signals fall within the bandwidth of the broadband amplifier, B_b. Their frequencies are f_{d1} and f_{d2}, both lying outside B_n. In a receiver with a linear pre-detector section these signals will be filtered out. In a receiver with a B-L-N structure combination frequencies $2f_{d1} - f_{d2}$, $3f_{d1} - 2f_{d2}$ and others can appear at the limiter output. They may fall within the passband of the narrowband amplifier, B_n, and, if they are strong enough, mask the signal. Thus, an increase in the B_b/B_n ratio improves the suppression of impulse noise, but there is an increase in the probability of the signal being masked by concentrated disturbances owing to an increase in their number in the passband of the broadband amplifier, B_b. Therefore, it is usual to choose B_b approximately equal to anywhere between 2.5 and 6 times the passband of the narrowband amplifier, B_n.

7.12. Diversity Reception

Diversity reception is an efficient method of reducing the effects of fading during the reception of a radio signal. This is done by using several instead of one realization representing the same intelligence. The rationale of the method is that it reduces the correlation between the realizations of the same intelligence, and so is the probability of the received signals fading all at the same time. The signals are then combined to enhance the fidelity and rate of message transmission. There may be space diversity, polarization diversity, and angle diversity.

In *space diversity*, the signal is received simultaneously by several antennas spaced a certain distance apart.

In *polarization diversity*, one utilizes receiving antennas which differ in polarization.

In *angle diversity*, one uses signals which arrive at the point of reception with some difference in the angle of arrival between the vertical and the horizontal planes.

In space-diversity reception, an antenna array of the smallest size is obtained when the receiving antennas are arranged to lie crosswise in a horizontal plane. Polarization diversity is especially good in minimizing the effects of polarization fading in the HF band; at microwave frequencies these effects are less troublesome. Angle diversity is most efficient at microwave frequencies.

Space-diversity reception is used most of all. It may use an array of two antennas (two-branch diversity), an array of three antennas (three-branch diversity), or an array of four antennas (four-branch diversity). Two-branch diversity is used most often. Three-branch diversity gives only a slight further improvement and is used more seldom. Four-branch diversity is a still rarer occurrence.

The received signal realizations are combined to form a combined signal

$$z_r(t) = \sum_{k=1}^{q} \varepsilon_k [\mu_k(t) x(t) + n(t)]$$

which is then processed by a decision device. In the above equation,

$\mu_k(t)$ = gain of the kth diversity branch
q = diversity order, that is to say, the number of signal realizations used to form the combined signal
ε_k = weight coefficient characterizing the method by which the combined signal is formed
$n(t)$ = noise

The combined signal can be formed by several combining methods, such as autoselection, linear or weighted combining, or both.

In autoselection, which may be non-optimum and optimum*, $z_r(t)$ is the realization of the received signal from the selected diversity branch. With all the other methods, the combined signal is formed by combining the signals from several branches. With non-optimum selection, the diversity branches, each containing an antenna and a receiver, are connected to a decision device which operates suitable switches in order to scan the branches in a fixed sequence and to select the one in which the signal-to-noise ratio, h_s^2, exceeds some specified threshold, $h_{s,th}^2$, that is the branch where the signal is the strongest. This branch is then used until h_s^2 falls below $h_{s,th}^2$, when it is disconnected and the scanning process starts again so as to select another branch for which $h_s^2 > h_{s,th}^2$, and so on. The weight coefficients in this form of diversity are

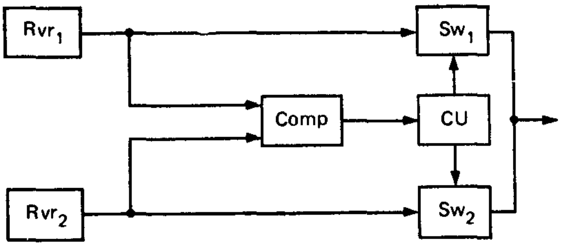


Fig. 7.7

$$\varepsilon_k = \begin{cases} 1 & \text{for } k = r \\ 0 & \text{for } k \neq r \end{cases} \qquad (7.7)$$

where r designates any branch in which h_s^2 is currently above the threshold, $h_{s,th}^2$. A limitation of this method is that the branch thus selected is just any one which satisfies the threshold condition rather than the one in which h_s^2 is a maximum. In contrast, with optimum autoselection, the selected branch is that for which $h_s^2 = h_{s,\max}^2$. The weight coefficients ε_k are defined by Eq. (7.7) as before, but r designates the best branch.

The scheme using autoselection and two-branch diversity is shown in Fig. 7.7. The output voltages of two receivers, Rvr_1 and Rvr_2, are applied to a comparator, *Comp*, and the difference signal goes to a control unit, *CU*, which operates two switches, Sw_1 and Sw_2. In linear combining, $\varepsilon_k = 1$, which means that both branches contribute

* Some authors call it 'scanning diversity' and 'selection diversity', respectively. See, for example W. C. Jakes, Jr., *Microwave Mobile Communications*. J. Wiley & Sons, 1974, p. 313 and p. 321.—*Translator's note.*

to the combined signal, irrespective of their h_s^2. For the branch signals to be combined with the same weight, all the branches should provide tbe same amount of amplification, which is ensured with the aid of a common AGC circuit*. The branch signals are cophased by an

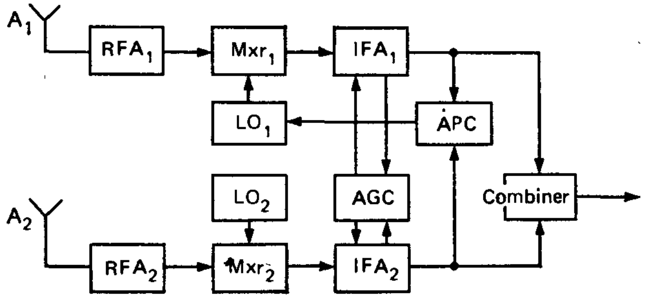


Fig. 7.8

automatic phase control circuit, APC (Fig. 7.8). In weighted combining**, $\varepsilon_k = (h_s^2)^{1/2}$. The branch signals are weighted in proportion to their signal-to-noise power ratios and then summed.Therefore, the branch with a deeper fade contributes less noise.

7.13. Adaptive Radio Links

In adaptive radio links the lack of prior knowledge about interference and noise is made up for by analysing the noise situation so as to obtain further information which could be used for optimal control of the receiver and transmitter. The radio resources available are then utilized most sparingly, but more time is spent to analyse and predict the channel status, and to find and implement an applicable decision.

The transmission of intelligence may be viewed as a process involving consecutive encoding and decoding of the signal $s(t)$ from the message source by the various elements of the radio link, namely, the transmitter, *Xmtr*, the transmission medium (the channel), and the receiver *Rvr*, each of which can be described by a functional operator of its own: L_1, L_2, and L_3 (Fig. 7.9). Then the signal present at the receiver output may be written as

$$s_2(t) = L_3 < L_2 \{L_1 [s_1(t)], \ y(t), \ n(t)\}$$

* Quite aptly, this method is known as *equal gain combining* outside the USSR. See W. C. Jakes, Jr., *Microwave Mobile Communications*. J. Wiley & Sons, 1974, p. 318—*Translator's note.*

** This is what is called *maximal ratio combining*. *Op.cit.*, p. 316.—*Translator's note.*

where $y(t)$ is an unintentional additive noise and $n(t)$ is the fluctuation noise.

Obviously, the functional operator L_2 does not lend itself to control. Therefore, the effects of the noise signals can only be made up for by a purposeful control of the operators L_1 and L_3. Optimum values of these two operators, $L_{1,0}$ and $L_{3,o}$, can be obtained by ways and means of varying complexity, depending on our knowledge about the operator L_2. Therefore, it is customary for an adaptive radio link to include a channel status analyzer, *CSA*, and a control unit (or units), *CU*, which provides for optimum control of receiver and transmitter parameters. Adaptive control is said to be effected around the small loop when it affects solely the receiver parameters, or around the large loop when it affects both the receiver and transmitter parameters.

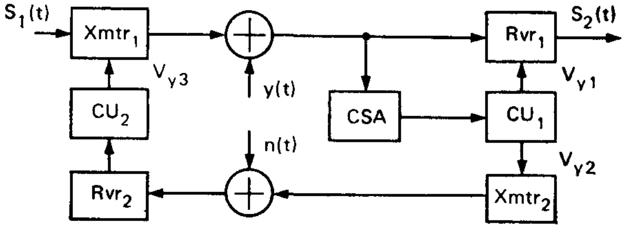


Fig. 7.9

Instead of being invariant with the noise situation, the error probability in an adaptive receiver is a function of the signal-to-noise ratio, h_s^2, and the functional error vector of the receiver subunits. This vector itself is a function of h_s^2, therefore with appreciable errors in realization an adaptive receiver may be inferior to a non-adaptive receiver in terms of noise immunity. The choice between an adaptive and a non-adaptive design depends on the purpose to be served by the system, the degree of uncertainty, and technological capabilities.

Communication channels in the HF band are subject to random variations in the signal level. Still, there are time intervals during which the signal level at some frequencies is sufficiently high. These frequency and time resources can be utilized in order to improve the reliability of communication by use of what are known as *frequency-adaptive radio links*. At each given instant, a frequency-adaptive radio link is using the frequencies which correspond to optimum condition. If all the working frequencies, f_W, allocated to a given radio link be uniformly distributed over the band, a panoramic receiver would be needed to analyse the channels for the presence and level of interference and noise and to locate open channels quickly. Some of these frequencies may prove unusable and more time would be required for a change-over from an unusable to a usable frequency. If, on the other hand, the working frequencies be grouped in compact form, the channel analysis will be simplified and less time will be needed for a frequency change, but frequency 'manoeuvrability' will be made more difficult and there will be a greater probability for the signals to be contaminated by common fades. That is why in frequency-adaptive radio links, the working frequencies are divided into groups, each group being such that $\Delta F_g \leqslant B$,.where B is the bandwidth of the preselector in the receiver and of the amplifier section in the transmitter. With this arrangement, there is no need to re-tune the transmitter and receiver when frequencies are changed within a given group, and the change itself is effected in the heterodyne (injection) frequency synthesizer (see Secs. 6.13 and 6.15). Since the conditions of wave propagation and the extent of channel contamination by noise and interference never remain constant with time, there is a need for a change of frequency groups. This means that in terms of organization a frequency-adaptive radio link is a feedback system similar to that shown in Fig. 7.9 where CU_2 does the job of a command and decision device which generates commands so that $Xmtr_1$ and Rvr_1 are retuned in synchronism.

Chapter Eight

Reception of AM Signals

8.1. Amplitude-Modulated Signals in the Transmission Medium

When designing any receiver, it is essential to provide for a high quality of message reproduction and noise immunity in reception. How well these requirements can be satisfied depends not only on the receiver design but also on the conditions in which radio waves propagate. In the transmission medium, signals may be mutilated so much that their correction in the receiver may prove difficult if at all possible. Furthermore, noise is superimposed on the signal as it propagates through the medium. The extent to which these effects can be minimized depends on antenna design, so in the general case one should treat an antenna and the associated receiver as an entity. In the case of mass-produced receivers for radio broadcasting and local communication, which use various, mostly very simple antennas, it is customary to limit the performance analysis to the receiver proper.

Despite the advent of novel radio communication systems, the overriding principle in the use of free-space radio wave propagation, which provides for electromagnetic compatibility of radio communication and broadcasting systems, is, as has always been, frequency allocation: each system is assigned a certain, relatively narrow band of frequencies. The principal forms of modulation are, as they have been so far, amplitude and angle modulation. The radio signals used in such systems have a relatively narrow band, so they may be treated as quasi-harmonic. In the general case, they may all be defined as

$$v_s(t) = V_s(t) \cos \psi_s(t) \tag{8.1}$$

where

$$\psi_s(t) = \omega_0 t + \xi_s(t)$$

whereas $V_s(t)$ and $\xi_s(t)$ are functions slowly varying in comparison with $\cos \omega_0 t$, and their frequency spectra are a small fraction of the carrier frequency, f_0.

As has been noted more than once, in order to make the results more definitive and also to simplify receiver testing for compliance with specifications, it is assumed for $V_s(t)$ and $\xi_s(t)$ in the case of continuous-wave signals that single-tone modulation applies. Then, for AM

$$V_s(t) = V_{s,0}(1 + m_s \cos \Omega_s t) \tag{8.2}$$

and for FM (see also Example 7.4)

$$\xi_s(t) = \xi_{s,m} \sin \Omega_s t \tag{8.3}$$

As a rule, the results thus obtained give a sufficiently accurate idea about the quality of the receiver. In some cases, however, a need may arise to use more elaborate signal models. In considering the reception of AM signals, let us deem that $\xi_{s,m} = 0$.

A feature most characteristic and significant in the real propagation of radio waves is multipath transmission, the phenomenon where the signals reach the receiving antenna over two or more paths which differ in length. Consider the effect of this phenomenon on radio reception, taking as an example the simplest case of two-path propagation.

When radio waves travel over long distances through the ionosphere, the signals reaching the receiving antenna over two path are shifted relative to each other by a time τ often equal to anywhere between 0.5 and 2 ms. Then the signal spectrum at the point of reception may be written as

$$\begin{aligned} v_s(t) = {} & V_{s0} \cos \omega_0 t + aV_{c0} \cos \omega_0 (t - \tau) \\ & + (m_s V_{s0}/2) \cos (\omega_0 - \Omega_s) t \\ & + a (m_s V_{s0}/2) \cos (\omega_0 - \Omega_s)(t - \tau) \\ & + (m_s V_{s0}/2) \cos (\omega_0 + \Omega_s) t \\ & + a (m_s V_{s0}/2) \cos (\omega_0 + \Omega_s)(t - \tau) \end{aligned} \tag{8.4}$$

where the coefficient a is, in the general case, other than unity, which means that the signals have the same amplitude in both beams.

It is seen that in two-path propagation the signal components at the angular carrier frequency ω_0 are shifted relative to each other in phase by an angle $\omega_0\tau$; the components with the side frequency $\omega_0 - \Omega_s$ are shifted relative to each other by an angle $(\omega_0 - \Omega_s)\tau$, and the components with the side frequency $\omega_0 + \Omega_s$ are shifted relative to each other by an angle $(\omega_0 + \Omega_s)\tau$. These shifts affect the envelope of the AM signal, which may, after detection, show up as signal corruption. The quality of reception is especially affected by the phase shift between the components at the carrier frequency.

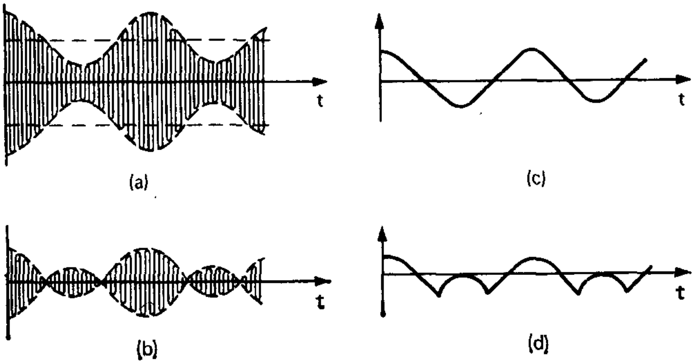


Fig. 8.1

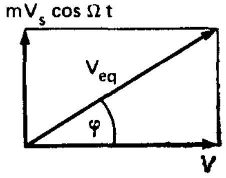


Fig. 8.2

When the phase shift $\omega_0\tau$ is approximately equal to π (180°), the components at the carrier frequency are in anti-phase, and the resultant amplitude falls (at $a = 1$, to zero). As will be recalled, an appreciable reduction in the carrier leads to overmodulation. This occurrence is illustrated in Fig. 8.1, where the plot at '*a*' shows an undistorted modulated signal, the plot at '*b*' shows a signal with a reduced carrier and unchanged side bands, whereas the plots at '*c*' and '*d*' show the signal at the output of the amplitude detector; it follows the envelope of the detected signal. As is seen from Fig. 8.1, the selective carrier fading which results from two-path propagation brings about heavy nonlinear distortion in the received message.

The actual picture is far more complex because the sideband components change in phase as well, a fact which upsets the equality of their amplitudes and leads to a further distortion. To prove, consider the limiting case of a complete carrier fade ($a \approx 1$, and $\omega_0\tau \approx \pi$). In such a case, the signal spectrum, defined by Eq. (8.4), will retain only the side frequencies $f_0 + F_s$ and $f_0 - F_s$. The resultant waves will be beats whose amplitudes vary at the difference frequency

$$(f_0 + F_s) - (f_0 - F_s) = 2F_s$$

irrespective of the phase angles of the components. In consequence, what will emerge from the amplitude detector will be an alternating voltage at frequency $2F_s$. That is, the sole result will be the 'distortion', and no message could then be received.

In the reception of real rather than single-tone-modulated signals, such as a telephone conversation with a complex spectrum, carrier fading will cause similar consequences. Deviations from the phase relationships in the sidebands and the selective fades of individual components which satisfy equalities of the form $(\omega_0 - \Omega_s)\,\tau \approx \pi$ or $(\omega_0 + \Omega_s)\,\tau \approx \pi$ will also cause distortion. This form of distortion is not so important because when one of the components in the upper side band fades, its counterpart in the lower sideband is usually preserved, and vice versa. Also the fades of some components in the sideband spectra do not lead to the destruction of the message as a whole.

Referring to Eq. (8.4), the summation of the carrier components with phases $\omega_0 t$ and $\omega_0\,(t - \tau)$ will not necessarily lead to fading (in fact, the amplitude may even increase), but the resultant wave will show a phase shift in any case, thus leading to heavy distortion. To demonstrate, suppose that the carrier phase has changed by $\pi/2$. Let us deem for simplicity that the sideband components have an epoch of zero, which means that the resultant wave consisting of the carrier and two side components has the form

$$v_s = V_s \sin \omega_0 t + (m_s V_s/2) \cos (\omega_0 + \Omega_s)$$
$$+ (m_s V_s/2) \cos (\omega_0 - \Omega_s)\, t$$

or, in a different form,

$$v_s = V_s \sin \omega_0 t + m_s V_s \cos \Omega_s t \cos \omega_0 t$$

It is seen from the phasor diagram of Fig. 8.2 that the amplitude of the resultant wave is

$$V_{s,eq} = [V_s^2 + (m_s V_s)^2 \cos^2 \Omega_s t]^{1/2}$$

or, in a different way,

$$V_{s,eq} = V_s\,(1 + 0.5\, m_s^2 + 0.5 m_s^2 \cos 2\, \Omega_s t)^{1/2}$$

On expanding the expression in the parentheses by the binomial theorem, limiting ourselves to the first terms of the expansion, and neglecting $0.5\, m_s^2$ in comparison with unity, we get

$$V_{s,eq} \approx V_s\,(1 + 0.25\, m_s^2 \cos 2\, \Omega_s t)$$

On top of all, the resultant voltage has a variable phase angle, φ, to which the amplitude detector does not respond.

The effect at the detector output is the same as in the detection of an AM signal with an angle modulation frequency $2\,\Omega_s$; instead of the transmitted message, the sole result will, as in the case of a complete carrier fade, be the 'distortion', and no message could then be received. Furthermore, the equivalent modulation factor at twice the modulation frequency is very small, being $m_{s,eq} \approx 0.25\, m_s^2$; if, for example, $m_s = 0.3$, then $m_{s,eq} \approx 0.02$. In consequence, the complete corruption of the transmitted signal is accompanied by a reduction in its modulation.

It is a fact, however, that intelligible reception is feasible in the HF band as well, although multipath interference is a common and strong occurrence there. The examples examined above hold for the worst conditions which may arise only occasionally. They prove that the quality of reception does not depend on the receiver alone. It also follows from the foregoing that distortion can be minimized through the use of highly directional antennas which are able to select preferentially one beam and to attenuate all the other arriving with a phase shift.

8.2. AM Signals in the Linear Section of a Receiver

Referring to Fig. 8.3*a* which shows an even-symmetric amplitude characteristic and an odd-symmetric phase characteristic of the pre-detector section of a receiver and the spectrum of an AM signal,

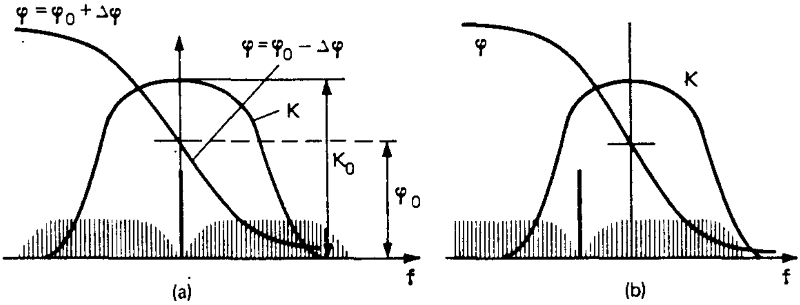


Fig. 8.3

it is seen that if the passband of the predector section is narrower than the signal spectrum width, the portion of the spectrum corresponding to the upper modulation frequencies will be attenuated. In consequence, that part of the spectrum will be attenuated at the detector output, which means that frequency distortion will take place.

Let us examine the effect of the phase characteristic, taking as an example the spectral components corresponding to some modulation frequency F_s. Together with the carrier, they make up a modulated wave of the form

$$v_s = V_{s0} \cos \omega_0 t + (m_s V_{s0}/2) \cos (\omega_0 - \Omega_s) t + (m_s V_{s0}/2) \cos (\omega_0 + \Omega_s) t \qquad (8.5)$$

Assuming that the gain K is the same for all the components, that is, neglecting frequency distortion, we obtain the following expression for the spectrum at the output of the pre-detector section:

$$v_{s,out} = KV_{s0} \cos (\omega_0 t + \varphi_0) + K (m_s V_{s0}/2) \cos [(\omega_0 - \Omega_s) t + \varphi_0 + \Delta\varphi] + K (m_s V_{s0}/2 \cos [(\omega_0 + \Omega_s) t + \varphi_0 - \Delta\varphi]$$

The above expression neglects the important fact that in a superhet the signal carrier at the output of the pre-detector section is different from what it is at the input.

It is an easy matter to re-cast the above expression as

$$v_{s,out} = kV_{s0} [1 + m_s \cos (\Omega_s t - \Delta\varphi)] \cos (\omega_{s0} t + \varphi_0)$$

Past a 'linear' amplitude detector of gain K_d, the alternating component of the signal will have the form

$$v = KK_d V_{s0} m_s \cos (\Omega_s t - \Delta\varphi)$$

Thus, the shape of the phase characteristic has an important bearing on the phase shifts between the spectral components of the signal at the detector output. If, for the receiver section in question, this response is practically a linear one, that is, if $\Delta\varphi = \tau\Omega_s$, then

$$v = KK_d V_{s0} m_s \cos \Omega_s (t - \tau)$$

which means that the signal solely suffers a group delay. If, on the other hand, the phase characteristic is not a linear one, phase distortion will arise.

The plot in Fig. 8.3*a*, represents the case when the receiver is exactly tuned to the desired signal. Actually, the receiver may be tuned other than exactly, which is often the case with simple broadcast receivers tuned manually 'by ear'. Given the same conditions as in the case illustrated in Fig. 8.3*a*, an example of inexact tuning is shown in Fig. 8.3*b*.

One of the consequences of inexact tuning, as can be seen from Fig. 8.3*b*, is the departure from proper phase relationship in the spectrum. In Sec. 8.1 it has been shown that this factor may be responsible for signal corruption. In the case at hand, changes in the phase shifts between the spectral components falling within the passband produce a relatively insignificant effect because within this band the phase characteristic usually only slightly differs from a linear one. The phase shift is about the same for all the spectral components, and the shape of the resultant AM wave does not change materially.

The quality of reception is more substantially affected by the following consequences of inexact tuning, seen from Fig. 8.3*b*: a reduction in the carrier amplitude if the carrier frequency lies on the sloping part of the response curve, and an attenuation of one sideband and a more complete transmission of the other.

As follows from Sec. 8.1, the reduction of the carrier may lead to heavy distortion. For this reason, a relatively large mistuning is intolerable because the carrier is then reduced to a marked extent. In the case illustrated in Fig. 8.3*b*, the carrier remains to lie within the passband, and the above event does not take place; it happens with a mistuning.

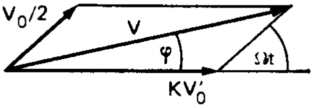


Fig. 8.4

It has been noted in Sec. 8.1 that some components of the sidebands may suffer selective fading due to multipath propagation. The case in Fig. 8.3*b* differs in that the greater part or all of one of the sidebands is attenuated.

In single-tone modulation, if one of the sidebands in the spectrum defined by Eq. (8.5) is completely suppressed, the AM signal has the form

$$v_{s,\mathrm{out}} = KV'_{s0} \cos \omega_{c0} t + K\,(m_s V_{s0}/2) \cos (\omega_{s0} + \Omega_s)\, t$$

where v'_{s0} is the carrier amplitude changed due to mistuning or for some other reason. The above expression does not take into account any additional phase shifts in the two components because they do not affect the final result.

As follows from the phasor diagram of Fig. 8.4, the amplitude of the total wave is

$$V_{\mathrm{out}} = K\,[(V'_{c0})^2 + (m_s V_{s0}/2)^2 + m_s V_s V'_{s0} \cos \Omega_s t]^{1/2}$$

The amplitude detector does not respond to the phase shift φ. The output voltage may be written as

$$V = KK_d V'_{s0} \left[1 + \left(\frac{m_s}{2} \frac{V_{s0}}{V'_{s0}} \right)^2 + m_s \frac{V_{s0}}{V'_{s0}} \cos \Omega_s t \right]^{1/2}$$

On expanding the above expression by the binomial theorem and limiting ourselves to the terms of the second order of smallness, we obtain

$$V \approx KK_d V'_{s0} \Big[1 + \frac{1}{8} m_s^2 (V_{s0}/V'_{s0})^2 + \frac{1}{2} m_s (V_{s0}/V'_{s0}) \cos \Omega_s t$$
$$- \frac{1}{16} m_s^2 (V_{s0}/V'_{s0})^2 - \frac{1}{16} m_s^2 (V_{s0}/V'_{s0})^2 \cos 2\Omega_s t \Big]$$

The alternating component of the detected signal has the form

$$V_{\mathrm{ac}} = KK_d \left[\frac{1}{2} m_s V_{s0} \cos \Omega_s t - \frac{1}{16} m_s^2 (V_{s0}^2/V'_{s0}) \cos 2\Omega_s t \right]$$

Thus, the reception is accompanied by nonlinear distortion, and the harmonic distortion factor is

$$k_h \approx (1/8)\, m_s V_{s0}/V'_{s0} \qquad (8.6)$$

Notably, when $V'_{s0} = V_{s0}$ and $m_s = 0.3$, the distortion factor is $k_h \approx 4\%$.

Nonlinear distortion increases with decreasing amplitude of the carrier, V'_{s0}. It can, therefore, be reduced by increasing V'_{s0}. This finding will be taken into account later, when we will be discussing single-sideband (SSB) reception.

Two diametrically opposite conclusions may be drawn from Fig. 8.3*b*:

(a) mistuning leads to an increase in nonlinear distortion, that is, to an impairment in message reproduction;

(b) mistuning broadens the frequency spectrum of the received message, that is, it causes a decrease in frequency distortion and, as a consequence, an improvement in the quality of reception.

Subjective tests show that if, in broadcast reception, the receiver bandwidth is made smaller than the signal spectrum, which is often done in order to improve selectivity (see Fig. 8.3*b*), an improvement in reproduction is ordinarily more significant than an increase in distortion. In such cases, inexact rather than exact tuning is preferable.

8.3. Distortion of the AM Signal in the Predetector Section of a Receiver

During the reception of strong signals, an adverse effect may take place due to the nonlinear volt-ampere characteristic of the electron devices used in the receiver. It is seen from Fig. 8.5 that when the amplitude of the input signal varies sinusoidally, the output signal may be distorted owing to the curvature of the amplitude characteristic.

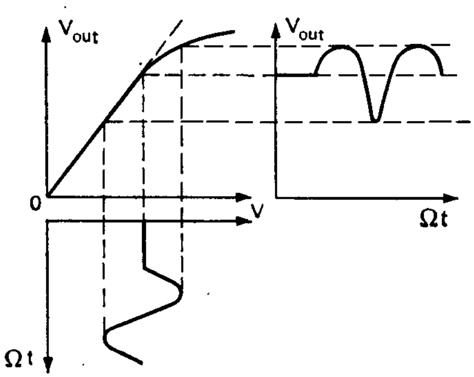


Fig. 8.5

Considering the portion of the curve where nonlinearity may show up, the amplitude characteristic may to a first approximation be written as

$$V_{s,\mathrm{out}} = K\,(V_s - \nu V_s^2)$$

where ν is a small empirical quantity and K is the small-signal gain.

When the modulation is as defined by Eq. (8.2), the amplitude of the output signal varies as

$$V_{s,out} = K\,[V_{s0}\,(1 + m_s \cos \Omega_s t) - \nu V_{s0}^2\,(1 + 2m_s \cos \Omega_s t$$
$$+ 0.5m_s^2 + 0.5m_s^2 \cos 2\,\Omega_s t)]$$

Thus, the output voltage of an amplitude detector contains the second harmonic whose amplitude is 0.5 $KK_d \nu m_s^2 V_{s0}^2$, and the harmonic distortion factor is

$$k_h = 0.5\, m_s \nu V_{s0}/(1 - 2\,\nu V_{s0})$$

In order to minimize distortion, it is essential to use electron devices whose characteristics have a broad linear portion and to avoid an excessive increase in the signal amplitude, V_{s0}.

8.4. Blocking Interference and Cross Modulation

In the linear section of a superhet, unless their spectra are superimposed on the spectrum of the desired signal, any interfering signals will be suppressed by the selective circuits of the i.f. amplifier. In a real amplifier, the events described in Sec. 8.3 occur in the final stages where the signal comes after it has been amplified. Consider the nonlinear effects that occur in the early stages of a receiver. In the early stages, the incoming signal is amplified very little, and their response to the signal is practically linear. Let us see what may happen if these stages have to handle both the desired signal and strong interfering signals from unwanted stations.

In contrast to the i.f. amplifier, the r.f. circuits mainly intended to reject the image and i.f. interference have a broad bandwidth. For example, while the passband of the i.f. amplifier of a broadcast receiver is several kilohertz, that of the input circuit and of the r.f. amplifier in the HF band is hundreds of kilohertz, and for the infradyne (see Sec. 4.5) it is several megahertz. Such a bandwidth encompasses interfering signals from tens or even hundreds of unwanted stations, and some of them, especially those coming from the nearest and high-power stations, may be very strong. The total interference voltage across the input circuit of a receiver is often as high as several hundred millivolts. With input voltages like that, the r.f. section may no longer be treated as a linear one.

Since interfering frequencies may greatly differ from the wanted signal frequency to which the receiver circuits are tuned, the electron devices may as greatly differ in behaviour, and this inevitably complicates their analysis. Let, to a first approximation, deem that the electron devices are free from inertia, which means that their properties are independent from frequency and the effect of the complex load on variation in the output current may be neglected.

Let us describe the usable portion of the volt-ampere characteristic of a nonlinear element by a Taylor series

$$i = \varphi(E + v) = \varphi(E) + \varphi'(E)\,v + (1/2!)\,\varphi''(E)v^2 + (1/3\ )\varphi'''(E)\,v^3 + \dots \qquad (8.7)$$

Here $\varphi(E) = I_0 =$ direct component of current at the operating (Q-) point
$\varphi'(E) = S =$ slope of the tangent to the curve at the operating point, or transconductance
$\varphi''(E) = S' =$ first derivative of the transconductance
$\varphi'''(E) = S'' =$ second derivative of the transconductance

Therefore, Eq. (8.7) may be re-cast as

$$i = I_0 + Sv + (S'/2)\,v^2 + (S''/6)v^3 + \dots \qquad (8.8)$$

A considerable rise in v usually slows down the rise in the current, which implies that S'' is a negative quantity.

When the wanted signal with a carrier frequency f_s, and an accompanying interfering signal with a carrier frequency f_{int}, are applied to an electron device, there appear at its output currents at the same frequencies, accompanied by harmonics along with sum and difference frequencies $kf_s \pm nf_{int}$. The succeeding filters tuned to the wanted signal frequency will pass only the components at frequency f_s.

Let the input voltage be a mixture of the wanted and an interfering signal

$$v = V_s \cos \omega_s t + V_{int} \cos \omega_{int} t \qquad (8.9)$$

On substituting Eq. (8.9) in Eq. (8.8) and separating (by simple manipulation of trigonometric functions) the current component at f_s, we find the amplitude of this component to be

$$I_s = SV_s + (1/8)\,S''V_s^3 + (1/4)\,S''V_sV_{int}^2 + \dots$$

In the case of strong interfering signals (such that $V_{int} \gg V_s$),

$$I_s \approx SV_s + 0.25\,S''V_sV_{int}^2 = I_{s0} + I_{s,int} \qquad (8.10)$$

where

$$I_{s0} = SV_s$$

is the signal current component in the absence of interference, and

$$I_{s,int} = 0.25\,S''V_sV_{int}^2$$

is the component current at the wanted signal frequency due to the effect of the interference.

It is seen from Eq. (8.10) that with $S'' < 0$, the signal level at the output of a nonlinear two-port decreases owing to a decrease in the mean transconductance under the action of the interference, this decrease being the greater, the greater the nonlinearity parameter (S'') and the interference amplitude (V_{int}). The decrease in the signal amplitude owing to interference is referred to as *blocking interference*. This effect is evaluated in terms of the blocking-interference factor

$$k_{\text{b.i.}} = I_{\text{s,int}}/I_{\text{s}0} = (1/4)\,(S''/S)\,V_{\text{int}}^2 \qquad (8.11)$$

Blocking interference can be mitigated by reducing the interference level with the aid of selective input circuits and by using electron devices which have a nearly linear characteristic, that is, those with a small ratio S''/S.

If the voltage reaching the input of a receiver is a mixture of a signal and a modulated interference whose amplitude is

$$V_{\text{int}}\,(t) = V_{\text{int}}\,(1 + m_{\text{int}} \cos \Omega_{\text{int}}t) \qquad (8.12)$$

and whose frequency, Ω_{int}, is other than the desired signal frequency and the spurious responses, the interference will be suppressed owing to the selectivity of the i.f. amplifier and will not reach the detector directly. However, the nonlinearity of electron devices is responsible for a process which leads to an irradicable distortion of the signal by the interference.

On substituting in Eq. (8.10) the signal amplitude defined by Eq. (8.2) and the interference amplitude defined by Eq. (8.12), and neglecting the relatively small terms, we get

$$I_{\text{s}} \approx SV_{\text{s}0} + Sm_{\text{s}}V_{\text{s}0} \cos \Omega_{\text{s}}t + 0.5\, S''V_{\text{s}0}V_{\text{int}0}^2 m_{\text{int}} \cos \Omega_{\text{int}}t + \dots$$

or, in a different way,

$$I_{\text{s}} = SV_{\text{s}0}\,[1 + m_{\text{s}} \cos \Omega_{\text{s}}t + 0.5\,(S''/S)m_{\text{int}}V_{\text{int}0}^2 \cos \Omega_{\text{int}}t + \dots] \qquad (8.13)$$

The third term in the above expression arises because the modulation of the interfering signal is transferred onto the desired signal carrier. This is known as *cross modulation* or *cross-talk interference*. Although the interference defined by Eq. (8.12), at frequency ω_{int}, does not pass through the selective section, its effect shows up after the detection has taken place because the detector output voltage is modulated by Ω_{int}. This effect is evaluated in terms of the cross-modulation factor, $k_{\text{c-m}}$, defined as the ratio between the parasitic modulation of the signal amplitude by the interference, which, according to Eq. (8.13), is $0.5\, m_{\text{int}}V_{\text{int}0}^2 S''/S$, and the modulation by the message being transmitted, m_{s}, that is,

$$k_{\text{c-m}} = (1/2)\,(m_{\text{int}}/m_{\text{s}})/S''/S)V_{\text{int}0}^2 \qquad (8.14)$$

It is seen from Eqs. (8.14) and (8.11) that cross modulation (or cross-talk interference) can be minimized in the same way as blocking interference (see also Sec. 7.5).

8.5. Intermodulation Interference

Suppose that the desired signal $V_s \cos \omega_s t$ arrives at the receiver input in company with interfering signals $V_{\text{int}1} \cos \omega_{\text{int}1} t$, $V_{\text{int}2} \cos \omega_{\text{int}2} t$, etc. The epochs of these alternating voltages are not taken into account because they will not affect the analysis that follows. By extending the series given in (8.8), the usable portion of the volt-ampere characteristic of the electron device to whose control electrode the above mixture is applied, may be approximated by a polynomial of the form

$$i = I_0 \sum_{k=1}^{n} a_k v^k$$

On substituting for v the above mixture of voltages and re-arranging the trigonometric functions, it is an easy matter to see that the spectrum of current i contains components at angular frequencies ω_s, $\omega_{\text{int}1}$, $\omega_{\text{int}2}$, . . ., and harmonics $2\omega_s$, $3\omega_s$, $2\omega_{\text{int}1}$, $3\omega_{\text{int}1}$, $2\omega_{\text{int}2}$, $3\omega_{\text{int}2}$, and combination (sum and difference) frequencies ($m\omega_{\text{int}1} \pm n\omega_{\text{int}2} \pm p\omega_{\text{int}3} \pm \ldots$). The components for which the sum $m + n + p + \ldots$ is a minimum will have the largest amplitudes. Since the receiver is tuned to ω_s, nearly all of these components will not fall within the bandwidth of the i.f. section after the frequency conversion has taken place and will not effect the reception of the desired signal. Some of the combination frequencies may, however, lie close enough to ω_s. They will fall within the bandwidth to be amplified along with the desired signals. They will be superimposed on the signal and will distort it.

The effects we have just described are referred to as *intermodulation* and the combination frequencies mentioned are often called intermodulation frequencies. The most troublesome intermodulation products are those of the third order, which arise due to the term $a_3 v^3$ of the approximating function. (As is seen from Eq. (8.8), $a_3 = S''/6$). A good deal of nuisance comes from the components with angular frequencies $2\omega_{\text{int}1} - \omega_{\text{int}2} \approx \omega_s$. In their case,

$$f_{\text{int}1} \approx (f_s + f_{\text{int}2})/2$$

which means that for such a component to be produced, it will suffice for the interfering frequency $f_{\text{int}1}$ to lie roughly midway between the signal frequency, f_s, and the second interfering frequency, $f_{\text{int}2}$.

Frequencies $f_{\text{int}1}$ and $f_{\text{int}2}$ may come from two high-power stations operating in adjacent or close-in channels. The spectra of such interfering signals and of the desired signal with frequency f_s are shown in Fig. 8.6. Here, curve *1* is the frequency response of the i.f. amplifier (assumed to be translated to f_s), and curve *2* is the response of the r.f. section. Neither $f_{\text{int}1}$ or $f_{\text{int}2}$ can pass through the receiver alone because the frequency produced by the frequency converter will lie outside the passband of the i.f. amplifier. Still, since they lie within the bandwidth of the r.f. section, they will produce in the electron devices of that section the intermodulation interference at frequency $2f_{\text{int}1} - f_{\text{int}2}$, which lies close to f_s and will, after the frequency conversion has taken place, fall within the passband of the i.f. amplifier along with the desired signal.

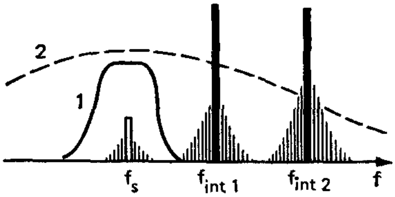


Fig. 8.6

The interfering frequencies which fall within the passband of the r.f. section (ahead of the frequency converter) may also give rise to components at frequencies $f_{\text{int}1} + f_{\text{int}2} - f_{\text{int}3} \approx f_s$, due to the same third-order, nonlinearity. With $f_{\text{int}1} \approx f_s + \Delta f$ and with $f_{\text{int}2} \approx f_s + 2\Delta f$ (where Δf may have any value), we have

$$f_{\text{int}3} \approx f_s + 3\Delta f$$

That is, the interfering signals are located in three equidistantly spaced frequency bands (say, in three adjacent channels).

If extraneous stations use amplitude modulation, the intermodulation interference they produce will likewise be amplitude-modulated by the messages they transmit. After detection, these interfering signals will be superimposed on the desired signal and will mutilate it so much that it may become unintelligible.

Similarly to cross-talk interference, intermodulation interference can be avoided by using highly linear input stages in the receiver and by protecting these stages against strong interference.

8.6. Detection of AM Signals in the Presence of an AM Interference

In practice it often happens that the i.f. amplifier output signal reaching the detector is accompanied by the valid signal of an unwanted station. Figure 8.7*a* illustrates a situation in which the detector is reached by an AM signal at frequency f_0 to which the receiver is tuned and by an AM interference at frequency f_{int} which falls within the receiver bandwidth, *B*. Figure 8.7*b* illustrates the case where f_{int} lies outside the receiver bandwidth, *B*, but it still reaches the detector because of poor selectance (see the 'tail' of the frequency response). Apart from reaching the detector directly, an interference may be produced in the input stages of the receiver from the signals transmitted by unwanted stations due to intermodulation (see Sec. 8.5).

The amplitude of the beats between the signals of amplitude V_s and the interference of amplitude V_{int} can be found from the phasor diagram in Fig. 8.8. Here

$$\Omega_b = \omega_{int} - \omega_s$$

is the angular beat frequency.

The detector does not respond to the phase shift φ; its output voltage is solely decided by the amplitude of the input voltage.

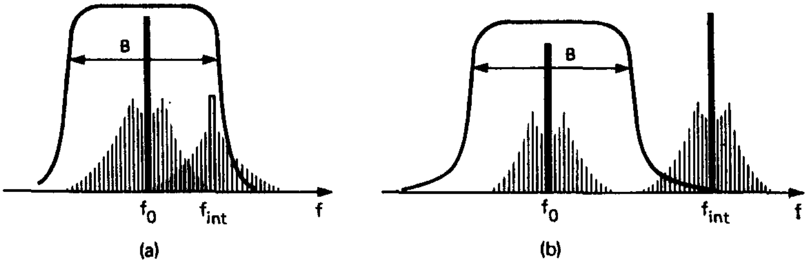


Fig. 8.7

Let us find this amplitude for $V_{int} < V_s$. The case of $V_{int} > V_s$ need not be analysed in detail because no satisfactory reception of messages will then be possible: the signal will be masked by the interference:

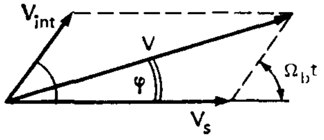


Fig. 8.8

As is seen from the phasor diagram,

$$V = (V_s^2 + V_{int}^2 + 2V_{int}V_s \cos \Omega_b t)^{1/2} \tag{8.15}$$

If $V_{int}^2 \ll V_s^2$, then

$$V \approx V_s [1 + 2 (V_{int}/V_s) \cos \Omega_b t]^{1/2}$$

On expanding the above expression by the binomial theorem and limiting ourselves to the first terms, we get

$$V \approx V_s (1 + m_b \cos \Omega_b t) \tag{8.16}$$

where $m_b = V_{int}/V_s$. This bears out the fact that beats are not unlike AM waves, with m_b being the equivalent modulation factor.

If the beats satisfy the condition defined in Eq. (5.35), the detector of gain K_d will produce an output voltage at the beat frequency

$$F_b = f_{int} - f_s$$

and of amplitude

$$V_{out} = K_d V_s m_b = K_d V_{int}$$

In the earphones or speaker of the receiver, one will then hear *whistles* or *tweets* at frequency F_b. That is how this form of interference, also called *heterodyne whistle*, predominantly shows up.

In most cases, the source of whistles or tweets is a station operating on an adjacent frequency channel so that the interfering frequency differs from that of the desired signal by a constant amount (often, by 9 kHz). The beat frequency shows a high stability because transmitters are usually built to stringent tolerances on frequency stability.

The frequency spectrum of the voltage appearing at the detector output in the reception of AM signals is shown in Fig. 8.9. Here, F_l and F_h are the lower and upper limiting frequencies of the transmitted spectrum, and F_b is the beat (or whistle) frequency.

Since F_b is highly constant, the interference can be cut out of the signal spectrum at the detector output by a narrowband rejection filter whose response is shown in the figure by the dashed curves. Of course, the filter will also cut out the signal frequencies falling within the rejection band, but if the rejection band is sufficiently narrow, the resultant distortion will be negligible. With $F_b > F_h$, the interference rejection will not produce any distortion at all. If F_b lies above the upper limit of audio frequencies ($F_b > 15$ kHz), the job of a rejection filter can be done by the human ear.

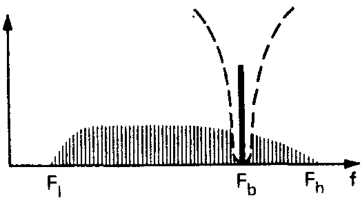


Fig. 8.9

After the interference at F_b has been suppressed, the spectrum of the detected signal will still contain components unaccounted for in the previous analysis and produced as the signal carrier beats with the sidebands of the undesired signal, but their intensity is lower. A still weaker effect is produced as the sidebands in the signal and interference spectra beat together.

Even if the interference at the beat frequency has been rejected by a filter, or attenuated owing to the detector's inertia, or is not perceived by the human ear, it may still show up at the modulation frequency as a result of direct detection. Consider two limiting cases:

(1) The detector is free from inertia, which means that the condition defined by Eq. (5.35) is satisfied as regards the beats.

(2) The condition defined in Eq. (5.35) is not satisfied, which means the detector load capacitor has no time to discharge during a beat period.

There may occur an intermediate case in which an analysis would yield likewise an intermediate result.

In case (1), the detector output voltage varies in proportion to variations in the voltage amplitude V.

Let us expand Eq. (8.15) into a power series and consider the terms of the next order of smallness in comparison with Eq. (8.16):

$$V \approx V_s\,(1 + 0.5\,\varepsilon - 0.125\,\varepsilon^2)$$

where

$$\varepsilon = (V_{\text{int}}/V_{\text{s}})^2 + 2\,(V_{\text{int}}/V_{\text{s}}) \cos \Omega_{\text{b}} t$$

On substituting for ε, discarding the components at the beat frequency and also the terms containing the ratio $(V_{\text{int}}/V_{\text{s}})$ to a power higher than second, and recalling that

$$\cos^2 \alpha = 0.5 + 0.5 \cos 2\alpha$$

we get

$$V \approx V_{\text{s}}\, [1 + 0.25\,(V_{\text{int}}/V_{\text{s}})^2 + \ldots]$$

On substituting for V_{int} from Eq. (8.12) and for V_{s} from Eq. (8.2), we get

$$V = V_{\text{s0}}\,(1 + m_{\text{s}} \cos \Omega_{\text{s}} t) + (1/4)\,(V^2_{\text{int0}}/V_{\text{s0}}) \times (1 + m_{\text{int}} \cos \Omega_{\text{int}} t)^2\, (1 + m_{\text{int}} \cos \Omega_{\text{int}} t)^{-1}$$

On expanding the factor in the last pair of parentheses by the binomial theorem and limiting ourselves to the alternating components at angular frequencies Ω_{s} and Ω_{int}, we get

$$V \approx V_{\text{s0}}\,(1 + m_{\text{s}} \cos \Omega_{\text{s}} t) + 0.5 V^2_{\text{int0}} V^{-1}_{\text{s0}} m_{\text{int}} \cos \Omega_{\text{int}} t$$

The a.c. voltage at the detector output, equal to $K_{\text{d}} V$, contains a component with frequency Ω_{s} and amplitude $V_{\text{s,out}} = K_{\text{d}} V_{\text{s}} m_{\text{s}}$,

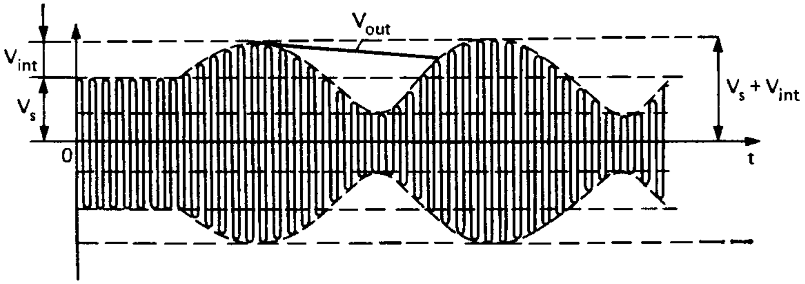


Fig. 8.10

and a component with frequency Ω_{int} and amplitude $V_{\text{int,out}} = K_{\text{d}} 0.5\,(V^2_{\text{int0}}/V_{\text{s0}}) m_{\text{int}}$.

Past the detector, the interference-signal ratio is

$$V_{\text{int,out}}/V_{\text{s,out}} = (1/2)\,(V_{\text{int0}}/V_{\text{s0}})^2 (m_{\text{int}}/m_{\text{s}}) \qquad (8.17)$$

Equation (8.17) shows that an amplitude detector has the property of amplitude selectivity. If V_{s0} exceeds V_{int0} by a factor of, say, 10, then, with $m_{\text{int}} \approx m_{\text{s}}$, the interference strength at the detector output will be 1/200th of the signal strength; obviously, it will then be hardly noticeable.

In rough terms, what happens in the second case can be gleaned from Fig. 8.10. After it has charged to a voltage approximately equal to the maximum beat amplitude, $V_s + V_{int}$, the detector load capacitor has no time to discharge; rather, it maintains V_{out} almost unchanged (shown by the heavy line) until the next peak. In view of Eqs. (8.2) and (8.12),

$$V_{out} \approx V_s + V_{int} = V_{s0}(1 + m_s \cos \Omega_s t) + V_{int0}(1 + m_{int} \cos \Omega_{int} t)$$

Hence,

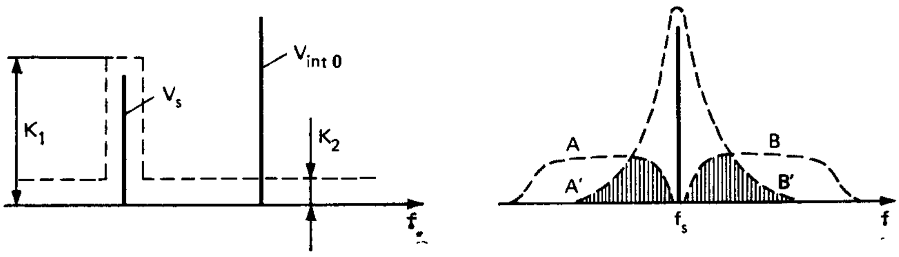
$$V_{int,out}/V_{s,out} = (V_{int0}/V_{s0})(m_{int}/m_s) \qquad (8.18)$$

That is, the detector reproduces both the interference and the signal without any change in their ratio, which means that the detector does not display its amplitude selectivity.

The inertia of the detector shows itself at a relatively high beat frequency, that is, when there is a marked difference between the interfering frequency and the signal frequency. In such cases, the interference is ordinarily attenuated to a sufficient degree by the predetector circuits of the receiver and does not show up, irrespective of the detector properties. The selectivity of the detector manifests itself when the interfering frequency lies close to the signal frequency, in which case the selectivity of the linear section of the receiver may in all probability prove inadequate. This, too, bears out the positive role played by the detector.

8.7. The Stenode Receiver and Synchronous Detection

As has been noted, whistles or tweets can be suppressed with relative ease if the carriers involved remain constant. What results from direct detection of the interfering signal can likewise be attenuated, but this calls for a more elaborate approach.



(a) (b)

Fig. 8.11

Consider the response of a receiver to a mixture of an AM signal of carrier frequency f_s and amplitude V_s, and an interference of carrier frequency f_{int} and amplitude V_{int}. The positions of the two carriers on the axis of frequencies is shown in Fig. 8.11a (the sidebands are omitted). In the case of practical interest, f_s and f_{int} lie close to each other. If, now, $V_s > V_{\text{int}}$, then Eq. (8.17) will hold true. In the example shown in Fig. 8.11 this inequality is not satisfied: the interference amplitude is about the same as the signal amplitude or even exceeds it. It follows then that Eq. (8.17) does not apply.

Suppose that the detector is preceded by a filtering network whose frequency response is shown in Fig. 8.11 by the dashed line. The gain of this network at frequency f_s is K_1, so that the signal carrier amplitude becomes equal to

$$V'_{s0} = K_1 V_{s0}$$

At any other frequencies of the spectrum, let the gain be equal to K_2. Then the interference carrier amplitude will be

$$V'_{\text{int}0} = K_2 V_{\text{int}0}$$

If K_2 is a small fraction of K_1, then $V'_{\text{int}0}$ will be smaller than V'_{c0}, and it is legitimate to use Eq. (8.17).

For the signal sidebands, the gain is likewise equal to K_2. Suppose that ahead of the filter the signal sidebands have an amplitude equal to $(m_s/V_{s0})/2$, where m_s is the modulation factor. At the filter output, these components have an amplitude equal to $(m_s V_{s0}/2)\,K_2$, which may be written as $m'_s V'_{s0}/2$, where

$$m'_s = m_s\,(K_2/K_1)$$

Thus, a change in the signal-to-sideband amplitude ratio has led to a change in the modulation factor. For the interference the modulation factor has remained unchanged because for the sidebands the gain (K_2) is the same, so that at the detector input $m'_{\text{int}} = m_{\text{int}}$.

The interference-to-signal ratio past the detector can be found, using Eq. (8.17) on replacing for $V'_{\text{int}0}$, V'_{s0}, m'_s and m'_{int}:

$$\begin{aligned} V_{\text{int,out}}/V_{\text{s,out}} &= 0.5\,(V'_{\text{int}0}/V'_{s0})^2\,(m'_{\text{int}}/m'_s) \\ &= 0.5\,(V_{\text{int}0}/V_{s0})^2\,(m_{\text{int}}/m_s)\,(K_2/K_1) \end{aligned}$$

The above result shows that the effect of the interference can be mitigated by making the ratio K_1/K_2 large enough. This arrangement provides a means for suppressing the interference even when $v_{\text{int}0}$ is greater than V_{s0}.

The frequency response shown dashed in Fig. 8.11*a* is not at all easy to implement. There was a receiver (developed and used in the 1930s) utilizing the above principle, in which the narrowband filter had an ordinary frequency response similar to that shown in Fig. 8.11*b*. Owing to its action, the signal sidebands were deformed. For example, if, ahead of the filter, the envelopes of the sidebands had the shape similar to curves *A* and *B*, then past the filter they took the shape represented by curves *A′* and *B′*. In other words, the components corresponding to the higher modulation frequencies were attenuated in comparison with the lower modulation frequencies (lying closer to the carrier). To avoid distortion in the audio-frequency (a.f.) amplifier, a frequency compensation circuit was included in the receiver known as the *stenode* in the history of radio engineering. Unfortunately, a number of difficulties stood in the way of its large-scale use, notably the inability of the narrowband filter to stay tuned to the signal carrier.

Figure 8.12 shows a device similar to the stenode in the operating principle but far more efficient. Here, *FSA* is a frequency-selective amplifier, *LPF* is a low-pass filter which passes the detected signal spectrum, and *LO* is the local oscillator whose voltage $V_{\rm LO}\cos\omega_s t$ is the same in frequency and phase as the incoming signal carrier $V_s\cos\omega_s t$.

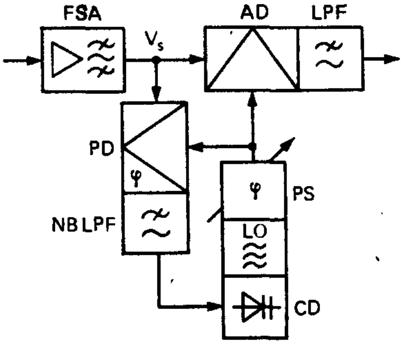


Fig. 8.12

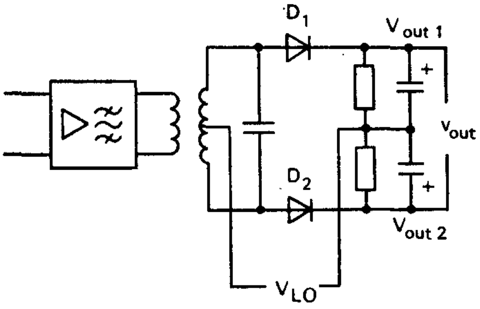


Fig. 8.13

In the detector the carrier is combined with the signal, $V_{s0}\,(1 + m_s\cos\Omega_s t)\cos\omega_s t$, and the total voltage takes, as it did in the previous case, the form given by $V'_{s0}\,(1 + m'_s\cos\Omega_s t)\cos\omega_s t$, such that

$$V'_{s0} = V_s + L_{\rm LO}$$

and

$$m'_s = m_s V_{s0}/(V_{\rm LO} + V_{s0})$$

but the voltage magnitude and modulation factor of the interference remain unchanged. As follows from Eq. (8.17), the interference-to-signal ratio past the detector is given by

$$\begin{aligned} V_{\rm int,out}/V_{s,\rm out} &= 0.5\,(V_{\rm int0}/V'_{s0})^2\,(m_{\rm int}/m'_s) \\ &= 0.5\,(V_{\rm int0}/V_{s0})^2\,(m_{\rm int}/m_s)\,[V_{s0}/(V_{s0} + V_{\rm LO})] \end{aligned}$$

It is seen that the effect of the interference can be moderated in a substantial way by increasing the local-oscillator voltage, $V_{\rm LO}$.

The above detection technique accompanied by a reduction in the interference is called *synchronous detection* (or *synchronous demodulation*), and the receiver incorporating a synchronous detector is called the *synchrodyne.*

As has been shown in Fig. 8.6, the intensity of whistles or tweets is solely proportional to the interference voltage amplitude. It follows then that neither the stenode nor the synchrodyne can avoid this form of interference in reception.

If the diode amplitude detector has the simple configuration shown in Fig. 5.13 and is sluggish in its response towards beats, neither the stenode nor the synchrodyne can suppress the interference. This can readily be proved on replacing V'_{int0} and m by V'_{int0} and m'_{int}, respectively, in Eq. (8.18). If a synchronous detector is to suppress the interference in this case as well, it must be set up in the balanced circuit of Fig. 8.13. In this arrangement, the voltage V_{LO} supplied by the synchronized local oscillator is applied to diodes D_1 and D_2 in phase, whereas the signal voltage V_s picked off the transformer secondary is applied in antiphase. The two voltages are chosen such that V_{LO} is substantially higher than V_s. As a result, the amplitude of the composite signal at frequency f_s across the upper diode is

$$V_{c0,u} = V_{\text{LO}} + V_s$$

and that across the lower diode is

$$V_{s0,l} = V_{\text{LO}} - V_s$$

As has been shown in Sec. 8.6, given an interference of amplitude V_{int} and an inertial detector, the detected voltage across the load of diode D_1 will be

$$V_{\text{out1}} \approx V_{s0,u} + V_{\text{int}}$$

whereas across the load of diode D_2 it will be

$$V_{\text{out2}} \approx V_{s0,l} + V_{\text{int}}$$

The resultant voltage at the detector output will be

$$V_{\text{out}} = V_{\text{out1}} - V_{\text{out2}}$$

On substituting for the component voltages, we obtain

$$V_{\text{out}} = 2V_s$$

Thus, only the signal will exist at the output of a synchronous detector, and there will be no interference present.

If a diode detector is free from inertia with regard to beats, Eq. (8.17) will hold true. It is an easy matter to show then that a balanced detector will not cancel the interference, but the latter is suppressed as efficiently as it is in an unbalanced synchronous detector.

In the arrangement of Fig. 8.12, the local oscillator is maintained synchronized with the signal carrier by a phase-locked loop which includes a phase detector, *PD*, and a narrowband low-pass filter, *NBLPF*. Owing to its narrow bandwidth, the output voltage of the low-pass filter is not practically affected by either the signal modulation or the presence of interference. The local-oscillator frequency is controlled by a control device (usually a varactor), *CD*. The phase shifter, *PS*, maintains the required phase relationship between the local-oscillator voltage and the signal carrier.

8.8. Impulse Noise Limiting in AM Receivers

Figure 8.14*a* shows how strong noise impulses, NI_1 and NI_2, affect the waveform of the signal voltage across the load resistor of a diode amplitude detector. This action can be minimized by limiting the voltage at the *A-A* level which corresponds to the maximum value of the detected signal. An example of a simple limiter is shown in Fig. 8.15. The detector output is coupled to a diode, *D*, which operates as a switch. The diode is rendered conducting (the switch is closed) by applying to its anode a positive voltage from a power

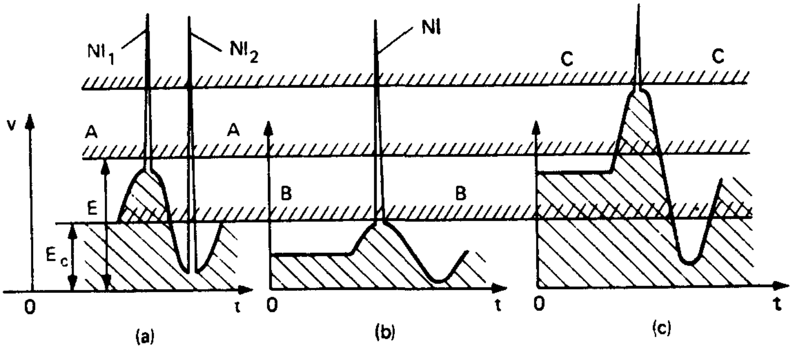


Fig. 8.14

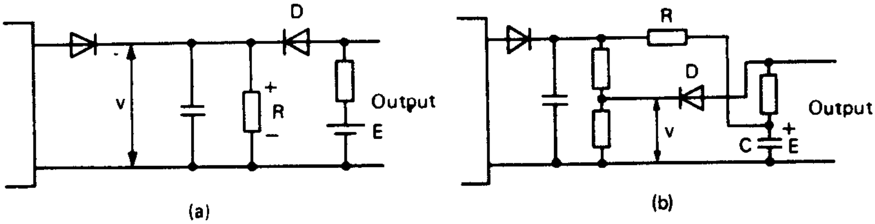


Fig. 8.15

supply having an emf (open-circuit voltage) E equal to the maximum voltage of the detected signal at a modulation factor of $m = 1$. The voltage picked off from the load resistor R is routed via the diode to the detector output.

When the voltage across the resistor R exceeds the emf E the resultant voltage at the diode anode becomes negative and turns off the diode (the switch is open). As long as the diode remains in the OFF condition, a constant voltage equal to E will be maintained at the circuit output. At $m = 1$, the maximum voltage at the detector output will be twice as great as the detected carrier voltage. Therefore, E is chosen to be approximately twice as great as the load voltage, $E \approx 2V_L$, which means that the limiting threshold is set at the A-A level.

If E were held always constant, the limiter would fail to operate as it should in response to variations in the signal voltage. If the signal decreased while E remained unchanged (Fig. 8.14*b*), the noise impulse, *NI*, passing through the limiter would substantially exceed the received signal. Therefore, in a case like that, it would seem preferable to set the limiting threshold at the B-B level. However, this level is unacceptable in the case shown in Fig. 8.14*a* because the limiter would then clip off both the noise impulse and some of the signal.

With an increase in the received signal voltage (Fig. 8.14*c*), the A-A and B-B thresholds would be unacceptable either, because the limiter would then clip off both the noise and the signal. It might seem reasonable to raise the limiting threshold to the C-C level, but it would be too high for the cases illustrated in Fig. 8.14*a* and *b*. Therefore, the limiter should be adjusted so that the limiting threshold goes up as the signal voltage rises and down as the signal voltage falls. This goal can be achieved by deriving the bucking voltage E from the incoming signal itself. An example of an impulse noise limiter with an adjustable threshold is shown in Fig. 8.15*b*. It does not differ from the previous one in the principle of operation, but the bucking voltage E is now derived from the detected signal by an RC low-pass filter. For this voltage to be twice the average value of v, the detector voltage reaching the load via the diode D is picked off the centre tap on the load resistor.

8.9. Reception of Double-Sideband Suppressed-Carrier Signals

The carrier of an AM signal does not contain any intelligence, still it uses or, rather, wastes, the greater part of the transmitter power. Therefore, no intelligence will be lost and a lot of power will be saved if the signal carrier is removed, or suppressed. A message signal in the form of two (upper and lower) sidebands without a carrier can be obtained by means of a balanced modulator. Then all of the transmitter power can be used to produce the two sidebands embodying the intelligence to be transmitted. This increase in the power of the useful part of the transmission spectrum minimizes the effect of additive noise on the quality of radio reception.

One of the techniques used in the reception of double-sideband suppressed-carrier (DSB-SC) signals consists in re-inserting the carrier in the spectrum of the signal at the receiver. This calls for a voltage of a suitable frequency and phase, which can be supplied by a local oscillator or a frequency synthesizer. Should, however, the carrier phase be set inaccurately, the received message will be heavily corrupted or it will not be received at all (see Sec. 8.1).

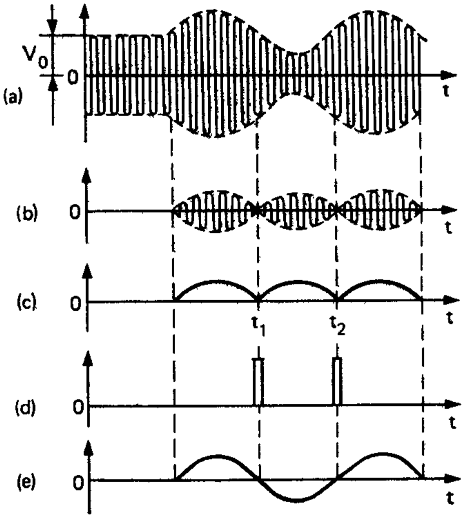


Fig. 8.16

For the carrier to be re-inserted at the receiver in the correct phase, it is usual to suppress it at the transmitter only in part so as to leave a few percent of its normal level still present. At the receiver, the residual carrier is extracted from the signal spectrum by a narrowband filter, amplified, and utilized by the phase-locked loop that "disciplines" the local carrier oscillator. This residual carrier used to control the frequency and phase of the reinserted carrier is called the *pilot carrier*, *pilot signal*, or simply the *pilot*. The voltage generated by the synchronized local oscillator and the pilot are added to the signal spectrum, that is, to its two sidebands. This action reconstructs a normal AM signal which is then detected by a conventional amplitude detector.

Because of its complexity, the above method of reception is not reliable enough. As has been stated in Sec. 6.11, a phase-locked loop can maintain the controlled oscillator so that its frequency is exactly the same as the signal frequency, but the two may differ in phase. Consider a different method which provides for a sufficiently exact detection in the absence of strong interference.

The waveform of a complete AM signal in which the carrier has an amplitude V_0 is shown in Fig. 8.16*a*. With the carrier suppressed, the signal takes the waveform shown in Fig. 8.16*b*. Amplitude detection produces a waveform shown in Fig. 8.16*c*. Its a.c. component contains only the second and higher harmonics and not a trace of the component at the modulation frequency. In order to convert this voltage to the transmitted message, a special circuit generates control pulses such as shown in Fig. 8.16*d* at instants when the voltage in Fig. 8.16*c* falls to zero (t_1, t_2, etc.). The circuit which serves to convert the rectified voltage is shown in Fig. 8.17. Here, *Sw* is a switch, and *CD* is a device which controls the switch (in practice, this is usually an electronic switch). Each pulse shown in Fig. 8.16*d* moves the switch from one position to the other. As a result, the voltage is reversed in polarity at times t_1, t_2, etc., and what is produced is a signal of the form shown in Fig. 8.16*e*, which is a faithful replica of the transmitted message.

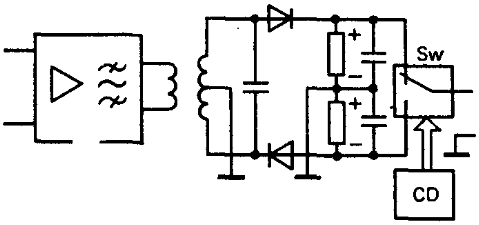


Fig. 8.17

8.10. Reception of Single-Sideband Signals

In Sec. 8.2 we have examined a case in which the receiver is not exactly 'on tune' and the detector is reached only by the carrier and one sideband of an AM signal. In this case, the reception is feasible and, indeed, more stable than with a double-sideband (DSB) signal because the phase shift of the carrier is of no importance. A drawback of this form of reception is an increase in nonlinear distortion, but, as follows from Eq. (8.6), the distortion can be reduced by artificially increasing the carrier amplitude, V'_{s0}.

The above form of signal is used in what is known as *single-sideband* (SSB) *transmission* and *reception*. As its name implies, the carrier and one sideband (either upper or lower) are suppressed at the sending end after the modulation has taken place. The remaining single sideband is then transmitted to the receiver. As follows from Sec. 8.9, given a transmitter of the same power rating, this permits a substantial increase in the power of the information-bearing emission and a reduction in the effect of interference and noise.

The carrier voltage, V'_{s0}, can readily be increased as the carrier is re-inserted at the receiver by means of a local oscillator or a frequency synthesizer. A frequency synthesizer will offer a simpler solution if the signal frequency is exactly known in advance.

Apart from a better utilization of transmitter power, single-sideband suppressed-carrier (SSB-SC) radio communication offers a number of other advantages, the most important among them being:

*Stability towards fading is improved. Signal fading may cause a reduction in the carrier amplitude or a change in the carrier phase. As has been noted, both factors contribute to the distortion of the modulated signal envelope and this shows up as a corruption of the received message after the demodulation has taken place.

*The required frequency spectrum is halved. This offers an opportunity to increase the handling capacity of a radio link by simultaneously sending two or more messages (for example, telephone conversations) from the same transmitter as independent SSB signals.

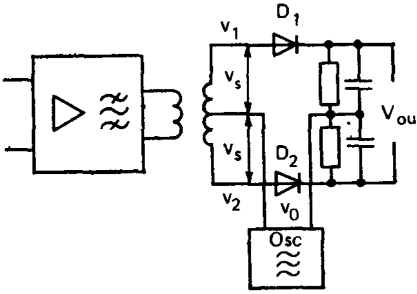


Fig. 8.18

Owing to their advantages, SSB systems have gained a wide field of application in radio communications and have ousted DSB systems with both an unsuppressed and suppressed carrier. Advances in microelectronics have made it possible to use SSB transmission in radio broadcasting as well, but at present this shift in policy would hardly be reasonable because millions of radio receivers not adapted to this form of operation are still in use throughout the world. Also, some progress has been made in stereophonic broadcasting at low and medium frequencies, which calls for modulation of an entirely different form (see Sec. 10.3).

The advent of SSB working has brought with it the need to solve three problems: a reduction in the nonlinear distortion caused by detection, generation at the receiver of a local carrier whose frequency is exactly equal to the original carrier suppressed at the transmitter, and automatic gain control.

Nonlinear distortion can be minimized by increasing the voltage supplied by the local carrier oscillator. As an alternative, use may be made of a balanced amplitude detector such as shown in Fig. 8.18. For its operation it depends on the well-known property of balanced circuits: attenuation of nonlinear distortion associated with even harmonics.

In the detector of Fig. 8.18, diode D_1 is fed a voltage V_1 equal to $V_{\text{LO}} + V_s$, and diode D_2 is fed a voltage V_2 equal to $V_{\text{LO}} - V_s$, where V_s is the SSB signal voltage picked off the centre tap on the transformer secondary. The SSB signal is assumed to be modelled by a voltage given by $V_s \cos(\omega_0 + \Omega t)$. The voltage supplied by the local carrier oscillator is

$$V_{\mathrm{LO}} = V_{\mathrm{LO}} \cos \omega_{\mathrm{LO}} t$$

Referring to a phasor diagram similar to that shown in Fig. 8.4, the amplitudes of the voltages across diodes D_1 and D_2 are found to be

$$V_1 = (V_{\mathrm{LO}}^2 + V_s^2 + 2V_{\mathrm{LO}}V_s \cos \Omega t)^{1/2}$$

$$V_2 = (V_{\mathrm{LO}}^2 + V_s^2 - 2V_{\mathrm{LO}}V_s \cos \Omega t)^{1/2}$$

Ordinarily, V_{LO}^2 is many times V_s^2, so the detector output voltage may be written as

$$V_{\mathrm{out}} \approx V_{\mathrm{LO}}K_{\mathrm{d}} \ \{[1 + 2\ (V_s/V_{\mathrm{LO}}) \cos \Omega t]^{1/2}, \\ - [1 - 2\ (V_s/V) \cos \Omega t]^{1/2}\}$$

Here, K_{d} is the detector gain. On expanding by the binomial theorem, we may re-write the above expression as

$$v_{\mathrm{out}} \approx V_{\mathrm{LO}}K_{\mathrm{d}} \ \{1 + (V_s/V_{\mathrm{LO}}) \cos \Omega t - (1/8)\ [2\ (V_s/V_{\mathrm{LO}}) \cos \Omega t]^2 \\ + \ \ldots -1 + (V_s/V_{\mathrm{LO}}) \cos \Omega t + (1/8)\ [2\ (V_s/V_{\mathrm{LO}}) \cos \Omega t]^2 \\ - \ldots\}$$

Since

$$\cos^2 \Omega t = 0.5 + 0.5 \cos 2\Omega t$$

it is inevitable that nonlinear distortion should take place in an unbalanced detector. At the output of a balanced detector, nonlinear distortion cancels out, and what is left is

$$v_{\mathrm{out}} \approx V_s K_{\mathrm{d}} \cos \Omega t$$

Odd-harmonic distortion associated with terms of higher powers will remain to be present, but at V_{LO}^2 many times V_s^2, it will be of minor importance.

In specifying the requirements for the accuracy of the locally re-inserted carrier it is important to remember that if it departs from its correct value, the signal frequency at the detector output will also be changed. If, instead of ω_{LO}, the angular frequency of the re-inserted carrier were $\omega_{\mathrm{LO}} + \Delta\omega$, the frequency of the voltage at the detector output would be

$$(\omega_{\mathrm{LO}} + \Omega_s) - (\omega_{\mathrm{LO}} + \Delta\omega) = \Omega_s - \Delta\omega$$

rather than Ω_s. It follows then that any departure of the locally inserted carrier oscillator from its correct value will corrupt the received message. Indeed, at $\Delta\omega = \Omega_s$, the message signal will not be reproduced at all because $\Omega_s - \Delta\omega = 0$.

For better insight into the consequences of a deviation of the locally re-inserted carrier oscillator from its correct value, consider the spectral components of a baseband signal at mutually multiple frequencies F, $2F$, $3F$, etc. Such baseband components are present in speech, music, etc. In a radio emission, these components correspond to frequencies $f_c + F$, $f_c + 2F$, $f_c + 3F$, where f_c is the carrier frequency. After detection in a receiver with a local carrier oscillator not exactly 'on-tune', these frequencies are converted to $F - \Delta f$, $2F - \Delta f$, $3F - \Delta f$, etc. It is easy to see that these frequencies are no longer multiples of one another. In reproduction, the sound will be hoarse and otherwise distorted. The distortion will be little noticeable if the local carrier oscillator is not over a few hertz 'off-tune'.

The requirements for the accuracy of the locally re-inserted carrier frequency are especially stringent in the case of high-fidelity music reproduction; here Δf must be not over 2 or 3 Hz. At $\Delta f \geqslant$ $\geqslant 20$ Hz, speech does not sound natural and recognizable, whereas at $\Delta f > 250$ Hz it becomes unintelligible.

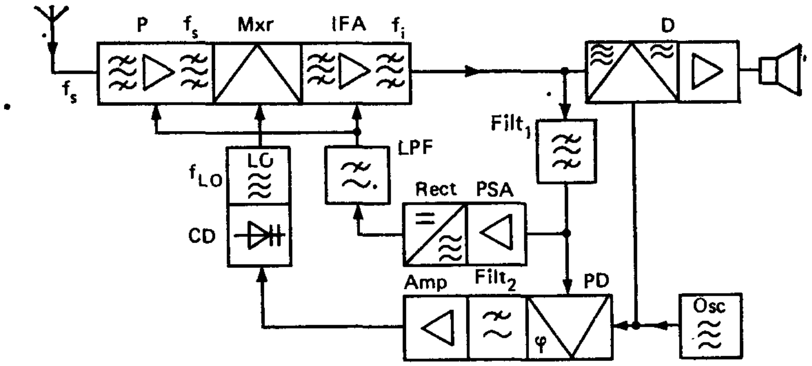


Fig. 8.19

Where radio links operate at fixed frequencies, the re-inserted carrier frequency is produced with sufficient accuracy and stability by a frequency synthesizer.

As has been noted, wide use is made of SSB systems with a pilot signal. In a receiver with a PLL, the pilot signal is the remainder of the carrier not completely suppressed in the transmitter. A simplified block diagram applying to such a case is shown in Fig. 8.19. Here, *P* is the preselector, *Mxr* is the mixer, *IFA* is the i.f. amplifier, and *D* is the detector.

The frequency of the carrier oscillator, *Osc*, is maintained exactly equal to the suppressed carrier by a phase-locked loop. For this purpose, the pilot signal separated by a narrowband filter, *Filt*$_1$, and the re-inserted carrier are applied to the inputs of a phase detector, *PD*, whose output voltage is smoothened by a second filter, *Filt*$_2$, and amplified, if necessary, by an amplifier, *Amp*, after which it is applied to the control device, *CD*, of the local oscillator, *LO*, used in the frequency converter.

A challenging task in SSB radio communications is automatic gain control which is an essential consideration for radio links subject to signal fading. In AM receivers, that is, those receiving DSB signals with an unsuppressed carrier, the control voltage for AGC is derived by rectifying the signal and passing it through a low-pass filter (see Secs. 6.3 through 6.5). The control voltage thus obtained is proportional to the carrier voltage. As the carrier decreases, the gain is increased, and vice versa. This kind of gain control cannot provide a constant voltage at the receiver output because changes in the carrier will not always be accompanied by fades in the sidebands whose voltage controls the output voltage of the detector. Still, it is utilized because it improves reception stability. In SSB working, the signal voltage varies not only because of fading, but to a greater degree according to the nature of the intelligence being transmitted. In speech transmission, for example, the transmitted sideband disappears during pauses and increases with increasing loudness. Because of this, the required control voltage cannot be derived by simple rectification of the signal as is done in the circuit of Fig. 6.3.

In the AGC circuits of SSB receivers with a pilot carrier, the control voltage is derived from the pilot signal after it has been rectified. This arrangement is shown in Fig. 8.19 where *PCA* is the pilot carrier amplifier, $Filt_1$ is the filter which extracts the pilot, *Rect* is the rectifier, and *LPF* is the low-pass filter. The voltage smoothened by the low-pass filter goes to the gain control circuit.

Gain control based on the pilot carrier level is similar to the AGC circuits of a DSB-AM receiver and suffers from the same drawbacks: variations in the pilot do not exactly follow variations in the received sideband level. If, instead of analog telephone signals with a randomly varying amplitude, a SSB link transmits signals with a fixed amplitude (as is the case with telegraph signals), the control voltage should be derived from the transmitted signals rather than from the pilot.

8.11. Reception of Digital AM Signals

Data generated by computer centres, digitized continuous signals and telegrams are transmitted by digital, mostly binary, radio links. In binary modulation systems, the digital information to be transmitted is assumed to be coded in binary form using two elementary signals. These two signals are called ‘mark’ and ‘space’ or ‘one’ and ‘zero’. These two signals can be generated by modulating, or *keying*, a sinusoidal carrier in amplitude, frequency, or phase in a time sequence of two mutually exclusive states. In *amplitude-shift* *keying* (ASK), the sinusoidal carrier is pulsed so that one of the binary states (usually a 'mark' or a 1) is represented by the presence of the carrier while the other state (a 'space' or a 0) is represented by its absence. The two signals are of equal duration T and occur with equal probability. As any other receiver, the one for digital signals consists basically of two sections: a frequency-selective amplifier section and a demodulator which detects and otherwise processes the incoming signals. A digital AM receiver differs from CW-AM receiver in how its demodulator is configured. A typical arrangement is shown in Fig. 8.20. The amplitude detector, AD, which follows the frequency-selective amplifier section, FSA, converts the input signal to d.c. pulses which are smoothened by a low-pass filter, LPF. The filter additionally filters out interfering signals if the receiver bandwidth ahead of the detector is greater than the optimal one. Past the low-pass filter, the smoothened pulses go to a threshold circuit, TC, where they are compared with a threshold level in order to make a decision as to the symbol transmitted (whether it is a 1 or a 0). Since the signal amplitude tends to vary owing to fading, the threshold level has to be made variable (so that it 'follows' the signal). This goal is achieved by a threshold-level detector, TLD, in which the time constant of the load is comparable with the correlation time of the signal fades, which means that it is greater than that of the main amplitude detector, AD. From the amplitude detector, the pulses go to the threshold circuit after some delay in the low-pass filter. This delay is provided in order that the threshold level can reach its steady-state value. Apart from decision-making, the threshold circuit shapes rectangular pulses. The effect of interfering signals during the space condition is minimized by a protection device (not shown in Fig. 8.20) which blocks the receiver during the space condition. The rectangular d.c. pulses appearing at the output of the threshold circuit can be used to drive the end unit directly, if the latter is located near the receiver. If, however, the end unit is located some distance from the receiver, audio-frequency (a.f.) pulses are fed into the connecting line (this serves to reduce the distortion of the pulses in the line). The required a.f. pulses are produced

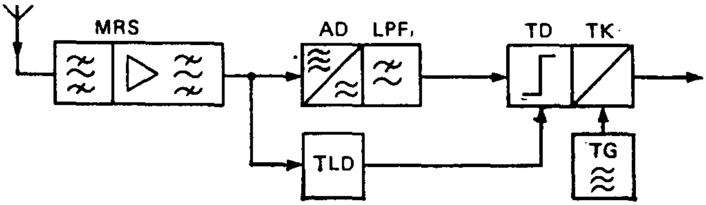


Fig. 8.20

by a tone keyer, *TK*, driven by a tone generator, *TG*. At the other end of the line, the tone signals are converted back to d.c. pulses by amplifiers-rectifiers and are then fed to the end unit.

Chapter Nine

Reception of Angle-Modulated Signals

9.1. Distortion of FM Signals due to Multipath Propagation

The signal model used in an analysis of a modulated-signal receiver can be described by Eqs. (8.1) and (8.3), noting that in the case of FM signals, V_s is constant. The modulation frequency F is ordinarily a small fraction of the centre frequency of an FM signal, f_0. Therefore, as in AM, it is assumed that the FM signal is a quasi-harmonic one, so it is legitimate to use a quasi-stationary method of analysis. Within each time slot which is a small portion of the modulation period, the signal may be treated as a harmonic one with an angular frequency

$$\omega(t) = d\psi_s(t)\, dt$$

Therefore,

$$\omega(t) = \omega_0 + d\xi_s(t)/dt = \omega_0 + \Delta\omega(t) \tag{9.1}$$

If the angular frequency deviation is

$$\Delta\omega = \Delta\omega_m \cos \Omega_s t$$

then

$$\xi_s(t) = \int\limits_0^t \Delta\omega_m \cos \Omega_s t\, dt = (\Delta\omega_m/\Omega_s) \sin \Omega_s t \tag{9.2}$$

In consequence, the amplitude of phase deviation in (8.3) is

$$\xi_{s,m} = \Delta\,\omega_m/\Omega_s = \Delta f_m/F_s \tag{9.3}$$

This quantity is called the *modulation index.*

Consider, as we did in Sec. 8.1 for AM, the two-path propagation of radio waves as a special case of multipath propagation. Again, similarly to the analysis of AM signals [see Eq. (8.4)], the signal at the point of reception in the present case may be written as

$$v(t) = V_s \cos[\omega_0 t + \xi_s(t)] + aV_s \cos[\omega_0 (t - \tau) - \xi_s(t - \tau)]$$

Assuming, as before, that a is less than unity, and referring to the phasor diagram of Fig. 9.1, the above expression may be re-cast as

$$v(t) = V(t) \cos[\omega_0 t + \xi_s(t) + \varphi(t)] \tag{9.4}$$

Here,

$$V(t) = [V_s^2 + (aV_s)^2 + 2aV_s^2 \cos\theta]^{1/2} \tag{9.5}$$

$$\varphi = \arctan[a \sin\theta/(1 + a\cos\theta)] \tag{9.6}$$

$$\theta = \xi_s(t - \tau) - \xi_s(t) - \omega_0\tau \tag{9.7}$$

The additional angle $\varphi(t)$ varies other than sinusoidally, that is, differently from the modulation of the signal for which the modulation index is $\xi_s(t)$.

In agreement with Eq. (9.4), the angular frequency of the resultant signal in the receiver antenna is

$$\omega'(t) = \omega_0 + \Delta\omega(t) + \delta\omega(t) \tag{9.8}$$

where

$$\delta\omega(t) = d\varphi(t)/dt$$

and ω' differs from ω defined in Eq. (9.1) by an additional term, $\delta\omega(t)$. Because $\varphi(t)$ varies other than sinusoidally, the additional modulation is likewise nonsinusoidal. Variations in the output voltage of the frequency detector follow variations in the frequency, for which reason multipath propagation causes the transmitted spectrum to acquire components with frequencies other than the modulation frequency, which is another way of saying that the signal suffers nonlinear distortion. The voltages of these components are proportional to $\delta\omega(t)$, and that of the wanted signal is proportional to $\Delta\omega(t)$.

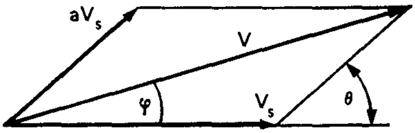


Fig. 9.1

The distortion decreases with decreasing amplitude of the delayed beam, that is, with decreasing a. When a is less than unity, which is usually the case when the two beams differ in amplitude, it is legitimate for a rough estimation of distortion to neglect $a\cos\theta$ in Eq. (9.6) in comparison with unity and to deem

$$\varphi \approx \arctan(a\sin\theta) \approx a\sin\theta$$

Then

$$\delta\omega(t) \approx a\,(d\theta/dt)\cos\theta$$

In view of Eqs. (9.7) and (9.2),

$$\delta\omega(t) \approx a\,\Delta\omega_m \cos\theta\,[\cos\Omega_s(t - \tau) - \cos\Omega_s t] \tag{9.9}$$

In broadband communication systems where F_s may be of the order of several megahertz or even tens of megahertz, the phase shift $\Omega_s\tau$ may be as great as π. Then $\delta\omega(t)$ may take on the maximum value at $\cos\theta \approx 1$ and $\cos\Omega_s(t - \tau) - \cos\Omega_s t \approx 2$. Then, $\delta\omega_m/\Delta\omega_m \approx 2a$.

Hence one can see the detrimental role played by delayed signals, such as those arriving at the point of reception over "dog-leg" paths after a number of reflections from mountains, buildings and other obstacles.

In many cases, such as in sound FM broadcasting in the VHF and UHF bands such that $F_{s,\max} \approx 10$ kHz, the phase shifts in the modulation of the delayed signals, $\Omega_s\tau$, are small. Noting that

$$\cos \Omega_s (t - \tau) - \cos \Omega_s t = 2 \sin (\Omega_s\tau/2) \sin \Omega_s (t - \tau/2)$$

and

$$\sin (\Omega_0\tau/2) \approx \Omega_s\tau/2$$

we may re-cast Eq. (9.9) as

$$\delta \omega (t) \approx a \, \Delta \omega_m \cos \theta \times \Omega_s\tau \sin \Omega_s (t - \tau/2) \qquad (9.10)$$

Here, in view of Eqs. (9.7) and (9.2),

$$\theta \approx (\Delta \omega_m/\Omega_s) [\sin \Omega_s (t - \tau) - \sin \Omega_s t] - \omega_0\tau$$

On re-writing the difference of sines in the usual manner and noting that $\sin \Omega_s\tau \approx \Omega_s\tau$, the above expression can readily be re-cast as

$$\theta \approx -\tau [\omega_0 + 2 \Delta \omega_m \cos \Omega_s (t - \tau/2)]$$

Re-writing the cosine of the sum of angles and noting that $\Delta \omega_m$ is small in comparison with ω_0, we may write $\cos \theta$ as

$$\cos \theta \approx \cos \omega_0\tau \cos [2\Delta \omega_m\tau \cos \Omega_s (t - \tau/2)]$$
$$- \sin \omega_0\tau \sin [2\Delta \omega_m\tau \cos \Omega_s (t - \tau/2)]$$

or, in view of the smallness of $\Delta \omega_m\tau$,

$$\cos \theta \approx \cos \omega_0\tau - 2\Delta \omega_m\tau \sin \omega_0\tau \cos \Omega_s (t - \tau/2) \qquad (9.11)$$

Substituting Eq. (9.11) in Eq. (9.10) gives

$$\delta \omega (t) \approx a\Delta \omega_m\Omega_s\tau \cos \omega_0\tau \sin \Omega_s (t - \tau/2)$$
$$- a (\Delta \omega_m\tau)^2\Omega_s \sin \omega_0\tau \sin 2\Omega_s (t - \tau/2)$$

Under the conditions in question, the first term in the above expression, $a \Delta \omega_m\Omega_s\tau \cos \omega_0\tau$, has a considerably lower amplitude than the main modulation component $\Delta \omega (t)$, equal to $\Delta \omega_m$, and it may therefore be neglected. The second component varies at twice the modulation frequency; its ratio to $\Delta \omega_m$ is the harmonic distortion factor, or the *harmonic content*. It is a maximum for $\sin \omega_0\tau = 1$, in which case,

$$k'_h \approx a \, \Delta{\cdot}\omega_m\Omega_s\tau^2$$

For example, if $a \approx 0.5$, $F_s \approx 5$ kHz, $\Delta f_m \approx 50$ kHz, and $\tau \approx 5\,\mu$s, then $k'_h \approx 12\%$. It is seen, therefore, that the nonlinear distortion caused by multipath propagation can be very strong, indeed.

Its occurrence has to be taken into account when choosing the type of and the location for receiving antennas. With a more directional antenna, it is less likely that the delayed signals picked up by the sidelobes can reach the receiver input.

Nonlinear distortion due to multipath propagation increases in magnitude in the case of an insufficient amplitude limiting during

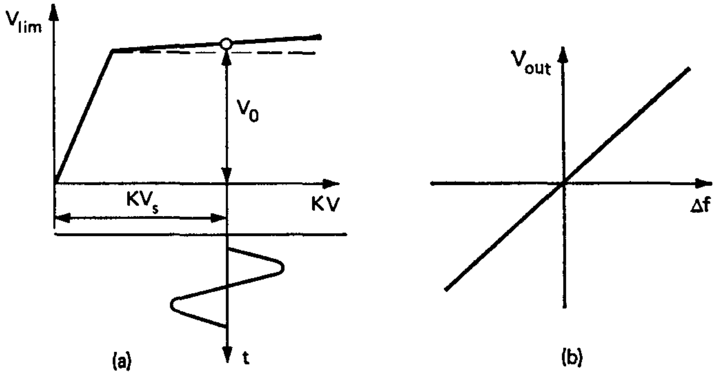


Fig. 9.2

the detection of FM signals. For an insight into the matter, suppose that the amplitude detector has a response of the form shown in Fig. 9.2*a* and the frequency discriminator, an ideal response of the form shown in Fig. 9.2*b*.

Let the voltage at the input to the limiter be equal to $KV(t)$, where K is the gain of the preceding stages, and $V(t)$ is given by Eq. (9.5). Assuming that $a^2 \ll 1$, replacing the square root by a power series, and limiting ourselves to the terms of the first order of smallness, we may re-write the above expression as

$$V(t) \approx V_s + \Delta V_s$$

where

$$\Delta V_s = aV_s \cos \theta$$

The voltage at the limiter output is

$$V_{\text{lim}} \approx KV_0 \left[1 + \mu \left(K \,\Delta V_s / KV_s\right)\right] \tag{9.12}$$

where μ is the slope of the limiter characteristic above the limiting threshold. For an ideal limiter, $\mu = 0$, whereas in the absence of limiting action, $\mu = 1$.

Let the voltage at the output of the frequency discriminator be equal to

$$V_{\text{out}} \approx V_{\text{lim}} K_{\text{FD}} \Delta f \tag{9.13}$$

where K_{FD} is a factor which takes care of the construction and circuit-element values of the frequency discriminator. If we neglect the frequency-modulation distortion identified earlier for purposes of an approximate analysis, then

$$\Delta f = \Delta f_m \cos \Omega_s t$$

Therefore,

$$V_{out} \approx KV_s (1 + \mu a \cos \theta) K_{FD} \Delta f_m \cos \Omega_s t$$

or, subject to Eq. (9.11),

$$V_{out} \approx KV_s \{1 + \mu a [\cos \omega_0 \tau - 2\Delta\omega_m \tau \sin \omega_0 \tau \cos \Omega_s \times (t - \tau/2)]\} K_{FD} \Delta f_m \cos \Omega_s t$$

The amplitude of its component at frequency Ω_s is

$$V\Omega_s \approx KV_s (1 + \mu a \cos \omega_0 \tau) K_{FD} \Delta f_m$$

Since

$$\cos \Omega_s (t - \tau/2) \cos \Omega_s t = (1/2) \cos (2\Omega_s t - \tau/2) + (1/2) \cos (\Omega_s \tau/2)$$

the output voltage also contains the second harmonic whose amplitude is

$$V_{2\Omega s} \approx KV_s \mu a \Delta\omega_m \tau \sin \omega_0 \tau K_{FD} \Delta f_m$$

Thus, the imperfection of the limiter is responsible for the fact that a further type of distortion is added to the above forms of distortion, for which the distortion factor is

$$k_h'' \approx \mu a \Delta\omega_m \tau \sin \omega_0 \tau/(1 + \mu a \cos \omega_0 \tau)$$

For $\sin \omega_0 \tau = 1$ and $\cos \omega_0 \tau = 0$,

$$k_h'' \approx \mu a \Delta\omega_m \tau$$

For example, if $\mu \approx 0.2$ and all the other conditions are the same as they were in the previous example, $k_h'' \approx 20\%$. This underscores the importance of having a good amplitude limiter: the distortion may be appreciable, but with μ tending to zero, k_h'' will also tend to zero.

9.2. FM Signals in the Linear Section of a Receiver

As under amplitude modulation (see Fig. 8.3), the FM signal passing through the pre-detector section of a receiver is jointly affected by the amplitude and phase characteristics of the section. In considering this effect, assume that the received signal has a constant amplitude V_s, and only its angular frequency is varying as

$$\Delta\omega = \Delta\omega_m \cos \Omega_s t$$

and so is its phase in accordance with Eq. (9.2). The assumption of a constant signal amplitude implies that the transmitter is treated as an ideal one, and the emitted waves propagate over a single path.

As the theory of signal transmission tells us, an FM signal has a frequency spectrum of a theoretically unlimited width. If the frequency deviation is Δf_m and the modulation frequency is F_s, the energy of the wave is practically concentrated within a bandwidth given by

$$B \approx 2\,(\Delta f_m + F_s)$$

Accordingly, the devices through which FM signals are to pass should have about the same passband. However, it is no less important that the device should have a uniform amplitude characteristic in the passband and a linear phase characteristic. To simplify the discussion, let us examine the effect of each of these two characteristics separately.

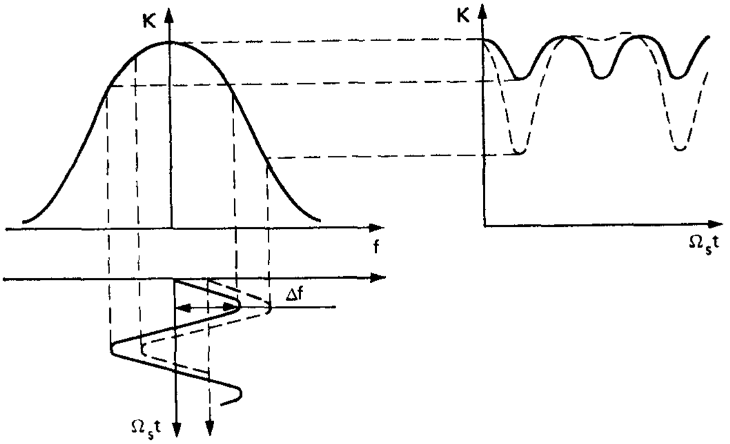


Fig. 9.3

If F_s is a small fraction of f_0, it may be deemed that the frequency is varying at relatively slow rate and that the gain K under modulation varies according to the static amplitude characteristic. The unbroken curve in Fig. 9.3 shows variations in K when the receiver is exactly tuned to the centre frequency, f_0, of the signal, whereas the dashed curve applies when the receiver is "off-tune". In the former case, K varies at a frequency equal to twice the modulation frequency, whereas in the latter case it tends to vary more at the modulation frequency, but the variations are not sinusoidal, that is, it contains the second harmonic. In each case, variations in K contain higher harmonics as well, but they will be neglected in our discussion for simplicity.

Suppose that K varies as

$$K \approx K_0 + K_1 \cos \Omega_s t - K_2 \cos 2\Omega_s t$$

Because of variations in the gain, a variable-frequency signal acquires a spurious amplitude modulation as it passes through the linear section of the receiver.

If the limiter is other than ideal, the output voltage of the frequency discriminator can be found in the same way as was done in Sec. 9.1, considering Eqs. (9.12) and (9.13) that is,

$$V_{\text{out}} \approx K_0 V_s \{1 + \mu\, [(K_1/K_0) \cos \Omega_s t - (K_2/K_0) \cos 2\Omega_s t]\}\, K_{\text{FD}}\, \Delta f_m \cos \Omega_s t$$

The fundamental component of V_{out} is

$$v_{\Omega s} \approx K_0 V_s K_{\text{FD}} \Delta f_m \cos \Omega_s t$$

Its second harmonic is

$$v_{2\Omega s} \approx (1/2)\, K_0 V_s \mu\, (K_1/K_0)\, K_{\text{FD}} \Delta f_m \cos 2\Omega_s t$$

and its third harmonic is

$$v_{3\Omega s} \approx -(1/2)\, K_0 V_s \mu\, (K_2/K_0)\, K_{\text{FD}} \Delta f_m \cos 3\Omega_s t$$

Hence, the total harmonic distortion is

$$k_h = (V_{2\Omega s}^2 + V_{3\Omega s}^2)^{1/2}/V_{\Omega s} = (\mu/2)\, [(K_1/K_0)^2 + (K_2/K_0)^2]^{1/2}$$

Thus, if the limiter is not efficient enough ($\mu \neq 0$), nonlinear distortion may arise not only due to multipath propagation, but also because of the spurious amplitude modulation arising in the receiver itself.

In considering the effect of the phase characteristic of an FM signal, assume, as before, that the signal arriving at the receiver input has the form defined by Eq. (9.1) and that the modulation is defined by Eq. (9.2). Since the modulation is a relatively slow process, we will, as we did before, use a quasi-stationary method of analysis. Accordingly, we assume that the output signal of the frequency-selective section of the receiver has the form defined by Eq. (9.4) where $\varphi\,(t)$ is the phase shift arising in that section.

The angle φ varies with time because it depends on the frequency which varies as defined by Eq. (9.2). We may therefore regard this angle more precisely as a function of the form $\varphi\, [\Delta\omega\, (t)]$.

To simplify the analysis, let us neglect variations in the amplitude V, that is, let us deem it to be constant, which can be achieved through the use of an effective amplitude limiter.

The phase characteristic of the linear section of an FM receiver is shown in Fig. 9.4. It differs from the plot in Fig. 8.3 solely in the scale: the quantity laid off on the axis of abscissas is the angular frequency. The dashed curve is the group delay, τ_d, of the signal, which is defined as the derivative $d\varphi/d\omega$.

The angular frequency of the signal at the output of the pre-detector section, ω', is similar to that defined by Eq. (9.8).

Also, as in Sec. 9.1,

$$\delta\omega(t) = d\varphi(t)/dt$$

which may preferably be written as

$$\delta\omega(t) = [d\varphi(\omega)/d\omega]\,(d\omega/dt) = \tau_d(\omega)\,(d\omega/dt)$$

The plot on the right of Fig. 9.4 shows variations in τ_d under modulation: the unbroken line applies when the receiver is exactly tuned to the centre frequency of the signal, and the dashed curve applies when it is 'off-tune'. If

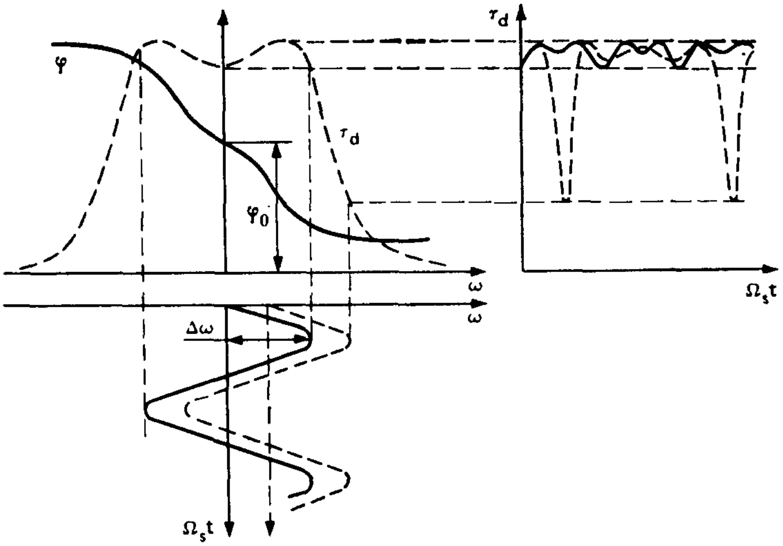


Fig. 9.4

$$\Delta\omega = \Delta\omega_m \cos\Omega_s t$$

then variations in the group delay may be expanded into a Fourier series as

$$\tau_d \approx \tau_0 + \sum_{k=1}^{n} \tau_k \cos k\Omega_s t$$

and, accordingly,

$$\omega(t) = d\psi_s(t)/dt = \omega_0 + \Delta\omega_m \cos\Omega_s t$$
$$+ \Delta\omega_m\Omega_s \sin\Omega_s t\left(\tau_0 + \sum_{k=1}^{n} \tau_k \cos k\Omega_s t\right)$$

On finding the output voltage of the frequency discriminator by Eq. (9.13) where $\Delta f = \Delta\omega/2\pi$ and recalling that in our case $V_{\text{lim}} =$

const, we get

$$v_{\text{out}} = (1/2\pi)\, V_{1m} K_{FD} \Delta\omega_m \{\cos \Omega_s t + \Omega_s \tau_0 \sin \Omega_s t + 0.5\Omega_s [\tau_2 \sin \omega_s t + (\tau_1 - \tau_3) \sin 2\Omega_s t + \tau_2 \sin 3\Omega_s t + \tau_3 \sin 4\Omega_s t + \ldots]\}$$

It is seen that the variations in modulation, related to the shape of the phase characteristic, show up after the detection (demodulation) has taken place in the form of nonlinear distortion. No distortion would have occurred if τ_1, τ_2, τ_3, etc. had been equal to zero. As follows from Fig. 9.4, this condition would have been satisfied if the group delay τ_d in the frequency swing $2\Delta f_m$ had had the form of a horizontal straight line, that is, if the phase characteristic in this frequency band had been linear.

9.3. Nonlinear Distortion in the Detection of FM Signals

Among the frequency detectors or discriminators used in FM receivers, those figuring most prominently are the types discussed in Sec. 5.10, where the FM signal is processed by resonant (tuned) circuits (see Figs. 5.32 and 5.35). Since amplitude detection takes place

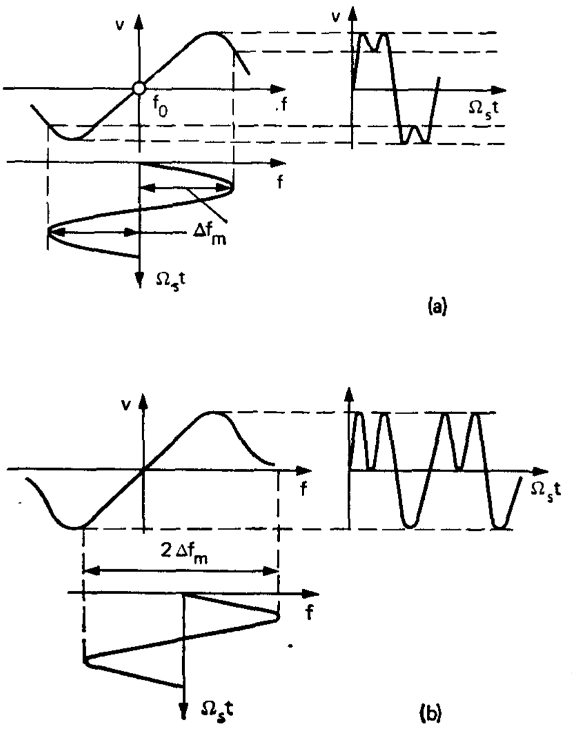


Fig. 9.5

in all of these cases, the nonlinear distortion examined in Sec. 5.5 (see Fig. 5.15) may well arise. It can be avoided, however, by satisfying the conditions defined there.

Nonlinear distortion may also be due to the curvature of the operating portion of the frequency detector's characteristic (see Figs. 5.36 and 5.39). This fact is taken care of by choosing the element values for the detector circuits so as to make the operating portion of the characteristic as linear as practicable.

Still other causes of nonlinear distortion may be the fact that the signal frequency swing, $2\Delta f_m$, is not equal to the width of the operating portion of the characteristic and that the receiver is not tuned exactly to the centre frequency of the signal. These two cases are illustrated in Fig. 9.5 from which it is seen that, even if the modulation is sinusoidal, the output voltage comes out distorted. In order to prevent an FM receiver from going 'off-tune' (Fig. 9.5*b*), it is usual to include an AFC loop.

9.4. Detection of FM Signals in the Presence of Noise and Interference

As with amplitude modulation, additive noise present under frequency modulation may distort the received message if the noise spectrum overlaps that of the signal in part or completely. An interference having an overlapping spectrum may also arise, as it does in the AM case (see Sec. 8.5), due to intermodulation.

Cross modulation may likewise occur in an FM receiver, but its mechanism is far more complicated than it is in an AM receiver (see Sec. 8.4). In some receivers, it is observed due to the breakthrough of an AM noise (this may be spurious AM, see Sec. 9.2) from the antenna via stray capacitances or other coupling elements to the local oscillator of the frequency converter. When a strong disturbance acts on the nonlinear interelectrode capacitances of the electron devices used in the local oscillator, it gives rise to a parasitic frequency modulation of the wave, and this modulation persists at the output of the frequency converter. As a result, the parasitic modulation is superimposed on the desired modulation of the signal and shows up after the frequency detection has taken place.

Consider how the reception of an FM signal is affected by a quasi-harmonic interference having a frequency different from the signal frequency. If the interference amplitude is smaller than that of the wanted signal (it is only then that a satisfactory reception will be possible), their addition can be illustrated by the phasor diagram of Fig. 8.8 where

$$\Omega_b = \omega_{int} - \omega_s$$

is the angular beat frequency. Owing to the superimposition of the interference, the signal suffers an additional phase shift given by

$$\varphi\,(t) = \text{arc tan}\,[V_{\text{int}} \sin \Omega_{\text{b}}t/(V_{\text{s}} + V_{\text{int}} \cos \Omega_{\text{b}}t)] \qquad (9.14)$$

In this analysis, we neglect variations in the frequency response of the receiver; rather, we assume that it has a flat gain, K. Then the signal applied to the frequency discriminator may be written as

$$v_{\text{out}} = KV\,(t) \cos\,[\omega_0 t + \xi_{\text{s}}\,(t) + \varphi\,(t)]$$

If the frequency discriminator provides good amplitude limiting, it may be assumed that the amplitude V is constant. Then the discriminator output voltage will be solely a function of frequency. Similarly to Eq. (9.8), the angular frequency of the signal may now be written as

$$\omega'\,(t) = \omega_0 + \Delta\omega\,(t) + \delta\omega\,(t)$$

where

$$\delta\omega\,(t) = d\varphi\,(t)/dt$$

The frequency-discriminator output voltage will contain the desired signal component proportional to

$$\Delta\omega\,(t) = \Delta\omega_m \cos \Omega_{\text{s}}t$$

and an interference component proportional to $\delta\omega\,(t)$. If we limit ourselves to a weak disturbance when the received signal is distorted but is still intelligible, we may neglect $V_{\text{int}} \cos \Omega_{\text{b}}t$ in the denominator of Eq. (9.14) and write $\varphi\,(t)$ as

$$\varphi\,(t) \approx (V_{\text{int}}/V_{\text{s}}) \sin \Omega_{\text{b}}t$$

Accordingly,

$$\delta\omega\,(t) \approx (V_{\text{int}}/V_{\text{s}})\,\Omega_{\text{b}} \cos \Omega_{\text{b}}t \qquad (9.15)$$

Thus, as is the case with AM, the interference shows up as whistles at the beat frequency superimposed on the detected signal.

The interference-to-signal amplitude ratio at the discriminator output is equal to the amplitude ratio of the frequency deviations $\delta\omega$ and $\Delta\omega$, that is

$$\begin{aligned} V_{\text{int,out}}/V_{\text{s,out}} &\approx (V_{\text{int}}/V_{\text{s}})\,(\Omega_{\text{b}}/\Delta\omega_m) \\ &= (V_{\text{int}}/V_{\text{s}})\,(F_{\text{b}}/\Delta f_{\text{m}}) \qquad (9.16) \end{aligned}$$

If, instead of one interfering signal with frequency f_{int}, several such signals with frequencies f_{int1}, f_{int2}, f_{int3}, etc. (to simplify matters, they may be assumed to have the same amplitude) reach the i.f. section, then each will manifest itself as explained above, but the voltages they produce at the frequency discriminator output will not be the same. As follows from Eq. (9.15), their effect will increase with an increase in the resultant beat frequency F_{b} or, which is the same, as f_{int} differs increasingly more from f_{s}. Then, at the output, the interference acquires the spectrum shown in Fig. 9.6*a*. The components of this spectrum are independent of whether the interfering frequency ($f_{\text{int}1}$, $f_{\text{int}2}$, or any other) lies above or below the signal

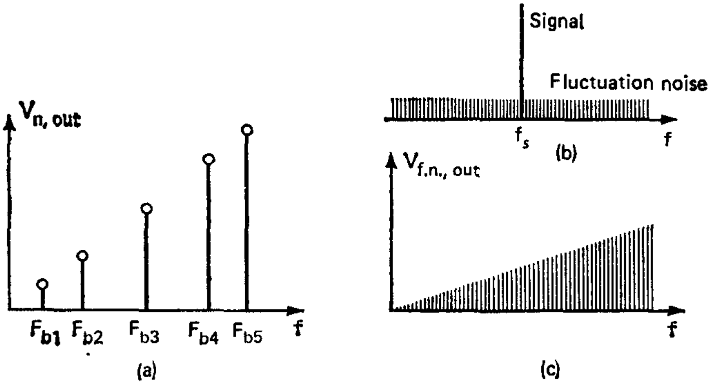


Fig. 9.6

frequency f_s. In some cases, the beat frequency will be equal to $f_s - f_{\text{int}}$, while in other cases it will be equal to $f_{\text{int}} - f_s$.

A similar picture emerges when the disturbance is fluctuation noise which has a continuous spectrum of a practically uniform density (see Fig. 9.6*b*). The spectrum at the output is shown in Fig. 9.6*e*.

Let us designate the product $V_{\text{lim}}K_{\text{FD}}$, which appeared in Eq. (9.13), as S_{FD}, that is, as the slope of the frequency discriminator (or detector) characteristic.

If $(\overline{V_{\text{n}}^2})_{\delta f_{\text{n}}}$ is the mean-square noise voltage in the band δf_{n}, then in our case (a uniform spectrum), the noise spectral density is

$$g = (\overline{V_{\text{n}}^2})_{\delta f_{\text{n}}}/\delta f_{\text{n}}$$

According to Eq. (9.15), the mean-square frequency deviation produced by the noise is

$$\overline{\delta f^2} \approx (g\delta f_{\text{n}}/V_{\text{s}}^2)F_{\text{b}}^2$$

The mean-square noise output voltage of a frequency detector can be found by integrating the components in the left- and right-hand halves of the spectrum in Fig. 9.6*b* with respect to the beat frequencies $F_{\text{b}} = f - f_{\text{s}}$ and $F_{\text{b}} = f_s - f$. Since the spectrum may be assumed to be symmetrical, it is legitimate to limit ourselves to only one half of the spectrum and then to double the result. In view of Eq. (9.15),

$$\overline{V_{\text{n, out}}^2} = S_{\text{FD}}^2 (g/V_{\text{s}}^2)^2 \int_0^{F_{\text{b, max}}} F_{\text{b}}^2\, dF_{\text{b}} = (2/3)\,(S_{\text{FD}}^2/V_{\text{s}}^2)\, F_{\text{b, max}}^2 g$$

Here, $F_{\mathrm{b,max}}$ is the maximum beat frequency at which the wave at the receiver output is perceived amidst the received message signal.

On the other hand, the effective (rms) output voltage associated with an FM signal is

$$V_{\mathrm{s,out}} = S_{\mathrm{FD}} \Delta f_m / 2$$

Therefore,

$$V_{\mathrm{n,out}}/V_{\mathrm{s,out}} = (2/\sqrt{3})\,(\sqrt{g}/V_s)\,(\sqrt{F^3_{\mathrm{b,max}}}/\Delta f_m) \qquad (9.17)$$

Both Eq. (9.16) and Eq. (9.17) show that the effect of interference can be mitigated by increasing the modulation, that is, by increasing the frequency deviation, Δf_{m}. Similar results are obtained under amplitude modulation where, according to Eq. (9.16), variations in the amplitude due to the beats produced by the superimposition of an interference on the signal are proportional to $m_{\mathrm{int}} = V_{\mathrm{int}}/V_{\mathrm{s}}$ (within the framework of our analysis, $m_{\mathrm{b}} < 1$), whereas the result of signal detection is proportional to m_{s}. There is, however, an important difference between the two cases: under AM, the modulation factor m_{s} can never be greater than $m_{\mathrm{s,max}} = 1$, whereas under FM, one can increase Δf_m without bound. Hence, we may conclude that broadband FM has a far better noise immunity than AM.

It is usually presumed that the maximum frequency of audible beats, $F_{\mathrm{b,max}}$, at which one should evaluate the interference-to-signal ratio by Eqs. (9.16) and (9.17), is the maximum modulation frequency for the signal, $F_{\mathrm{s,max}}$. This is a natural requirement if $F_{\mathrm{s,max}}$ is the frequency threshold of audibility for the human ear. If, however, $F_{\mathrm{s,max}}$ lies substantially below that threshold, the frequency discriminator should be followed by a low-pass filter with a cutoff frequency equal to $F_{\mathrm{s,max}}$, so as to suppress the higher interfering frequencies. Assuming a quasi-harmonic interference and considering Eq. (9.3), we then obtain for the two cases, respectively

$$V_{\mathrm{int,out}}/V_{\mathrm{s,out}} \approx (V_{\mathrm{int}}/V_{\mathrm{s}})\,\xi_{\mathrm{s},m} \qquad (9.16a)$$

$$V_{\mathrm{n,out}}/V_{\mathrm{s,out}} \approx (2/\sqrt{3})\,(\sqrt{gF_{\mathrm{s,max}}}/V_{\mathrm{s}})/\xi_{\mathrm{s},m} \qquad (9.17a)$$

where $\xi_{\mathrm{s},m}$ is the modulation index as found at $F_{\mathrm{s,max}}$.

Given the bandwidth B, the product gB is the mean square noise input voltage, $\overline{V_n^2}$. Therefore, Eq. (9.17a) may be recast as

$$V_{\mathrm{n,out}}/V_{\mathrm{s,out}} \approx (2/\sqrt{3})\,(V_{\mathrm{n}}/V_{\mathrm{s}})\sqrt{F_{\mathrm{s,max}}/B}\,(1/\xi_{\mathrm{s},m})$$

Since the bandwidth is proportional to the frequency deviation, the ratio $B/F_{\mathrm{s,max}}$ is proportional to the modulation index, $\xi_{\mathrm{s},m}$. It follows then that the ratio $V_{\mathrm{n,out}}/V_{\mathrm{s,out}}$ decreases in proportion to $\xi_{\mathrm{s},m}$, and the signal-to-noise power ratio increases in proportion to $\xi_{\mathrm{s},m}$.

In order to keep the interference to a sufficiently low level, it is usual to choose $\xi_{\mathrm{s},m} \approx 5$ to 7. In sound broadcasting, if the maximum modulation frequency is around 10 kHz, then the frequency deviation, Δf_m, ranges between 50 and 70 kHz. The width of the frequency band occupied by the emission in such a case is around 150 kHz. Such bands can only be allocated in the VHF and higher frequency bands where FM broadcasting is used.

FM receivers use the broadband amplifier-limiter-narrowband amplifier (B-L-N) structure discussed in Sec. 7.11. This structure is effective in suppressing impulse noise. In the pre-detector section, since it has a large bandwidth, an impulse noise signal is of short duration, but it has an appreciable amplitude which may be many times the amplitude of the wanted signal. In the narrowband filter which follows the frequency discriminator containing a limiter, the impulse. noise is stretched in time and decreased in amplitude so that it becomes weaker than the wanted signal.

In FM, the corrupting action of interfering signals is different at different frequencies of the received message. As is seen from Fig. 9.6*a* and *c*, the interference level is markedly lower at the lower frequencies and is a maximum at the upper frequencies. Unless measures are taken to the contrary, the quality of sound reproduction in FM transmission can heavily be impaired by the fact that the interference at the upper frequencies gains in strength. In order to provide approximately the same noise immunity at all modulation frequencies, it is customary in sound broadcasting to subject the transmitted spectrum to what is called *pre-emphasis*, that is, a process designed to emphasize the magnitude of some of the frequency components. For this purpose, the a.f. amplifier is arranged to have a frequency response such that the gain increases approximately in proportion to the increase in frequency. For this reason, the frequency deviation under modulation is likewise proportional to F_s. As a result, the spectrum of the detected signal is not unlike the noise spectrum at the detector output (see Fig. 9.6*c*), and the signal-to-noise ratio is therefore about the same and sufficiently high at any frequency in the band.

It is seen from Eqs. (9.2) and (9.3) that in transmission with pre-emphasis, the change in phase, that is the frequency modulation index $\xi_{c,m}$, is practically a constant quantity. In fact, this means that in such a case we have phase rather than frequency modulation.

An excessive boost at the high frequencies and an excessive cut at the lower frequencies would show up at the receiving end as severe frequency distortion. To avoid it, the frequency discriminator is followed by a frequency compensation network whose gain is inversely proportional to frequency. Owing to the action of this network, the transmitted spectrum is restored to its original form. The compensation network changes the gain to the same extent for both the signal and the noise, so the signal-to-noise ratio is left unchanged (see also Sec. 5.11, and the compensation network R_4C_4 in Fig. 5.40).

The foregoing and the original phasor diagram in Fig. 8.8 apply if the interference is weaker than the signal ($V_{\text{int}} \ll V_s$). This also holds true for a composite interfering signal if it is treated as a superposition of small interfering signals each of which, taken separately, satisfies the above condition. The overall effect of random small disturbances is illustrated in Fig. 9.7*a*. At any instant, they may be depicted by an overall phasor, V_{int}, that is, replaced by a single quasi-harmonic interference. The overall interference phasor V_{int}, drawn touching the tip of the phasor V_s which represents the signal, may make any angle with the latter, for which reason the phase shift may take on any value within certain limits. This is separately illustrated in Fig. 9.7*b* where the circle is the locus of points where the tip of the V_{int} phasor can be located; points A and B on this circle mark the limits between which the angle φ can vary. As long as $V_{\text{int}} < V_s$, the phase angle will be such that $\varphi < | \pm 90° |$, and

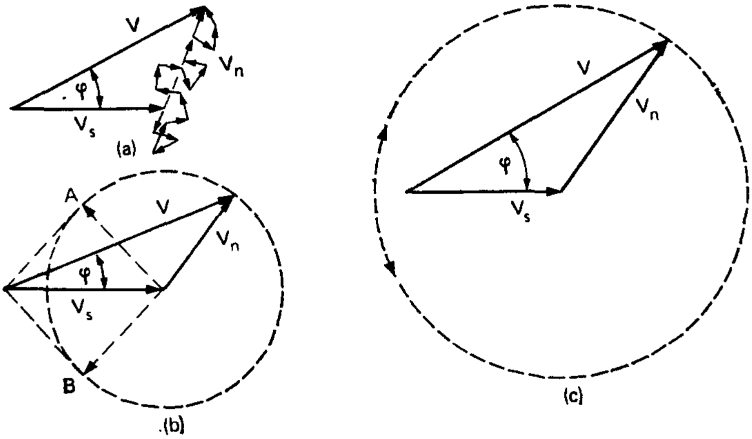


Fig. 9.7

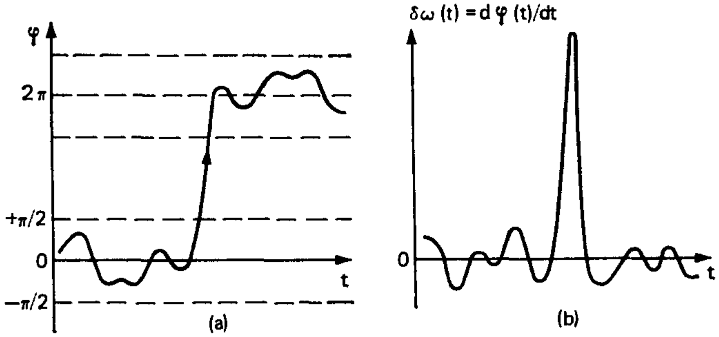


Fig. 9.8

it is only when V_{int} comes closer to V_s in magnitude that this angle is 90° very nearly. This explains why broadband angle modulation provides for high noise immunity if the modulation index is $\xi_s > \pi/2$.

Figure 9.7*a* shows that the V_{int} phasor may be random as well as the phase angle φ. With some probability, this phasor may exceed V_s at some instants. It is seen from Fig. 9.7*c* that this immediately removes the above constraint on the phase angle φ: it is now possible for the V phasor to rotate within a complete circle. An example of a change in the phase angle φ is shown in Fig. 9.8*a*, and an example of a change in the derivative $d\varphi/dt$, that is, the parasitic deviation of the angular frequency $\delta\omega$, produced by the superimposition of the interference on the signal, is shown in Fig. 9.8*b*. This change is the shape of an impulse and is responsible for the appearance of a voltage pulse at the output of the frequency discriminator.

To sum up, in angle modulation, the interference has a well-defined threshold character. High noise immunity is preserved only so long as $V_{\text{int}} < V_s$.

9.5. Threshold Reduction in FM Receivers

The lower the signal-to-noise ratio of a receiver, the more frequent it is that the noise amplitude spikes can exceed the signal level and the more frequent it is that the above-threshold impulse noise described earlier (see Fig. 9.8*b*) can show up. An overall picture of what happens in an FM receiver in the presence of fluctuation noise is shown in Fig. 9.9. Here, $h^2_{s,\text{in}}$ is the signal-to-noise power ratio ahead of the discriminator, and $h^2_{s,\text{out}}$ is the signal-to-noise power ratio at the discriminator output. It is seen that when $h^2_{s,\text{in}}$ is sufficiently high, the effect of the noise is markedly reduced, this reduction being the greater, the greater the modulation index $\xi_{c,m}$, which is fully in keeping with Eq. (9.16). A reduction in $h^2_{s,\text{in}}$ produces a threshold which separates the entire operating region into two qualitatively distinct regions: the above-threshold region and the below-threshold region. The onset of the threshold condition is accompanied by a drastic reduction in noise immunity. For example, with $\xi_{s,m} = 10$, the decrease in $h^2_{s,\text{in}}$ from 10 to 0 dB entails a decrease of about 30 dB in $h^2_{s,\text{out}}$. The threshold effect complicates the reception of FM signals

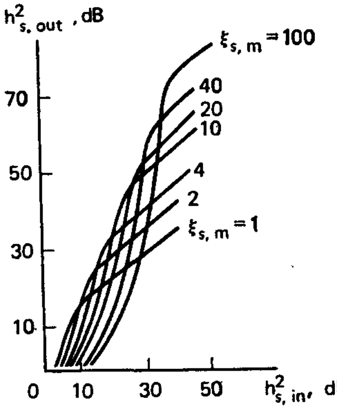


Fig. 9.9

in fading channels, and also in channels with a low energy potential close to the threshold one.

For the threshold to appear it is not mandatory for the average noise power to exceed the signal power. As $\xi_{s,m}$ increases, the operating frequency band is extended, and the noise spikes comparable in magnitude with the signal amplitude become more frequent, and this leads to a rapid fall in $h^2_{s,\text{out}}$.

The noise immunity of FM systems in the above-threshold region can be enhanced by the use of pre-emphasis described earlier. In order to suppress the threshold, or click, noise without having to increase the power output of the transmitter, resort is made to various threshold-reducing devices among which one of the most efficient is the tracking filter.

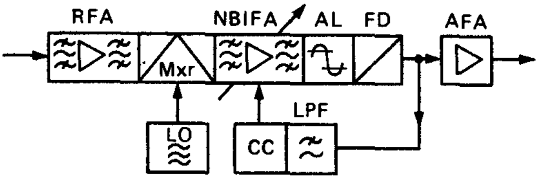


Fig. 9.10

The operation of a low-threshold FM receiver using a tracking filter consists in that one of the stages in the i.f. amplifier is arranged so that its bandwidth is somewhat narrower than the signal spectrum, and its resonant frequency is made to track the instantaneous frequency of the FM signal. The reduced bandwidth serves to minimize the noise voltage, whereas the signal voltage remains unchanged, provided the resonant frequency of the filter precisely tracks the signal frequency. As a result, the signal-to-noise ratio is markedly improved, and the threshold value is reduced. As is seen from Fig. 9.10, the narrowband i.f. amplifier, NBIFA, is made to track the signal frequency by a control circuit, *CC*. There is a low-pass filter, *LPF*, which makes this circuit insensitive towards the random fast changes in voltage that may be caused by the noise.

As an alternative, an FM receiver may use a voltage-controlled oscillator, *VCO*, and a feedback loop similar in its action to the AFC loop discussed in detail in Sec. 6.10. The arrangement of an FM feedback receiver is shown in Fig. 9.11. As the VCO output is injected in the mixer, the modulation of the intermediate frequency is reduced, and this permits the replacement of a broadband i.f. amplifier by a narrowband one. The fact that the feedback loop contains two narrowband filters, one for the i.f. and the other for the a.f., serves to bring down the threshold value.

Figure 9.12 shows an FM receiver using a PLL, where the job of a narrowband frequency converter is done by a phase detector, *PD*, followed by a low-pass filter, *LPF*. As in the previous case, the modulator is a voltage-controlled oscillator, *VCO*. The threshold is reduced owing to the relatively narrow passband of the low-pass filter and a proportionate reduction in the noise level. The detected signal emerging from the phase detector is routed via the low-pass filter to an audio-frequency amplifier, *AFA*, which means that the AFC loop also doubles as a frequency discriminator.

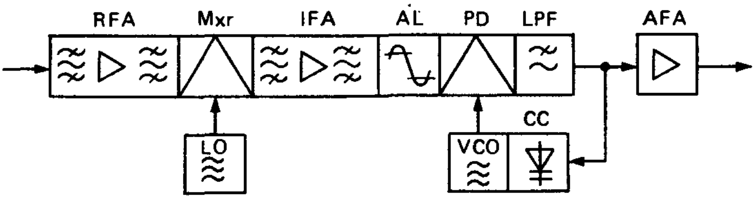


Fig. 9.11

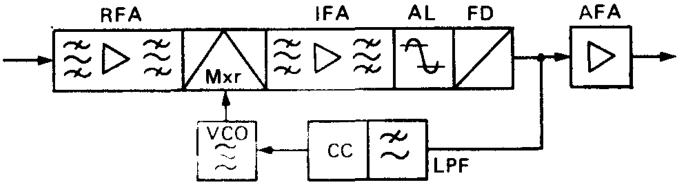


Fig. 9.12

9.6. Reception of Digital Messages in FM Systems

In a digital FM system, a 0 symbol is transmitted at frequency f_1, and a 1 symbol, at frequency f_2. Thus, the transmitter carrier is shifted back and forth in frequency in accord with the 1 and 0 intelligence of the binary code. Quite aptly, this form of modulation is known as *frequency-shift keying*. At present, it is the main type of radio transmission for digital messages. It owes its wider popularity in comparison with AM due to its better noise immunity in reception and the relative ease with which the incoming signal can be detected at the receiver. A simplified block diagram of a binary FM receiver is shown in Fig. 9.13. The signal picked up by the antenna is routed via the main reception section, *MRS*, to an amplitude limiter, *AL*, and then to a frequency discriminator, *FD*. The limiter eliminates changes caused in the signal amplitude by noise and fading. The detector produces bipolar (plus-and-minus) d.c. pulses which go through a low-pass filter, *LPF*, and reach a threshold device, *TD*. The pulses can be converted to tone signals by a tone-signal converter, *TSC*, before they are injected into a connecting line.

FM signals can be detected by any one of the frequency detectors or discriminators examined in Secs. 5.10 and 5.11. Wide use is made of the frequency detector shown in Fig. 9.14 which is similar to the double-tuned frequency detector of Fig. 5.32, except that instead of simple tuned circuits the amplifier section, *Amp*, has at its output

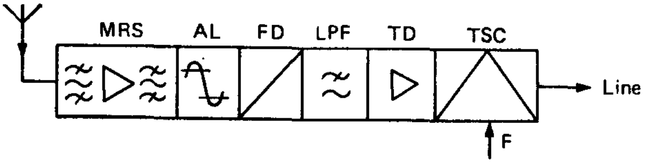


Fig. 9.13

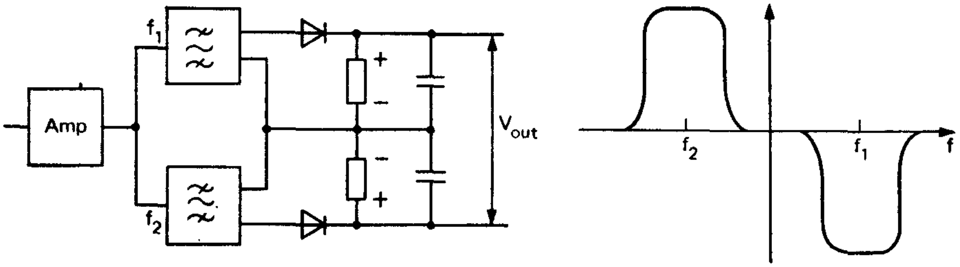


Fig. 9.14

more elaborate filters, including crystal units, for a better separation of the two frequencies, f_1 and f_2. The passband of each filter should provide for the separation of the main portion of the signal spectrum around frequencies f_1 and f_2 and is chosen according to the pulse duration or, which is the same, the keying rate. Thus arranged, the receiver has the B-L-N structure examined in Sec. 7.11, which serves to minimize impulse noise.

Use is also made of digital detectors, for example, of the *pulse-counting type*. They operate by counting the number of signal periods or cycles during the time T when an elementary signal is being transmitted. The decision as to whether a 1 (a "mark") or a 0 (a "space") has been transmitted is made by the threshold device which is fed the voltage from the output of the low-pass filter (see Fig. 5.44) or the integrator (see Fig. 5.46). In other detectors, the counter counts zero crossings (see, for example, Fig. 5.45) and gives the number n of half-periods or half-cycles. The decision is made by comparing the counted number n with the threshold value equal to $(n_1 + n_2)/2$, where n_1 and n_1 are the *apriori* known numbers of half-cycles of frequencies f_1 and f_2 over the interval T. This arrangement is shown in Fig. 9.15.

Still another type of digital frequency detector operates by measuring the duration of signal half-cycles. Its simplified arrangement is shown in Fig. 9.16. To begin with, the incoming signal is clipped at the top and bottom to give it a square shape by an amplitude gate, *AG*. The half-cycle duration is measured by counting the number of clock pulses occurring during that interval. The accuracy of measurement depends on the frequency of the clock generator, *CG*, or, rather, the repetition rate of clock pulses. For example, if a time interval is to be measured accurate to within 1%, the repetition rate of clock pulses must bẹ 100 times the signal frequency at the input to the frequency detector. The count is transferred to a digital-to-analog converter, *DAC*, and its analog output voltage is passed through a low-pass filter, *LPF*, to a threshold circuit, *TC*. The noise immunity of the detector can be enhanced if, in making the decision as to whether a given half-cycle is associated with f_1 or f_2, one takes into account the result obtained by measuring the duration of a previous and a succeeding half-cycle.

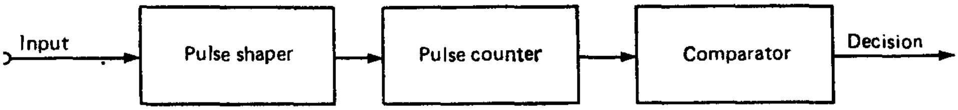


Fig. 9.15

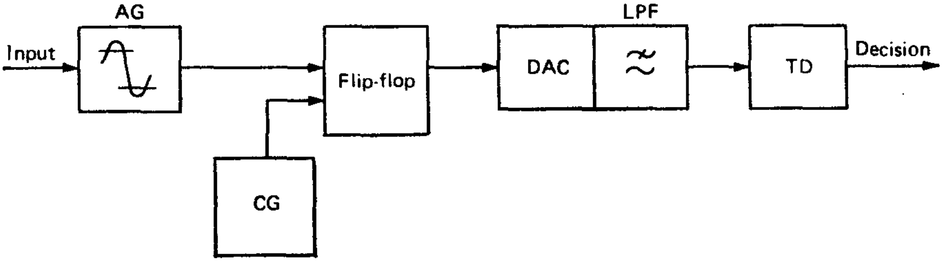


Fig. 9.16

Point-to-point radio links widely use double-channel frequency-shift keying, DC-FSK. In order to transmit four signal combinations over the two channels, one uses four frequencies: f_1, f_2, f_3, and f_4. For example, f_1 can be used to transmit a binary 0 over both channels, f_2 to transmit a binary 1 likewise over both channels, f_3 to transmit a binary 1 over the first and a binary 0 over the second channel, and, finally, f_4 to transmit a binary 0 over the first and a binary 1 over the second channel. With this arrangement, all of the transmitter power is utilized to produce signals in both channels. This provides for a maximum signal level and, as a consequence, a maximum noise immunity. An arrangement that can be used to separate the channels in a DC-FSK receiver is shown in Fig. 9.17.

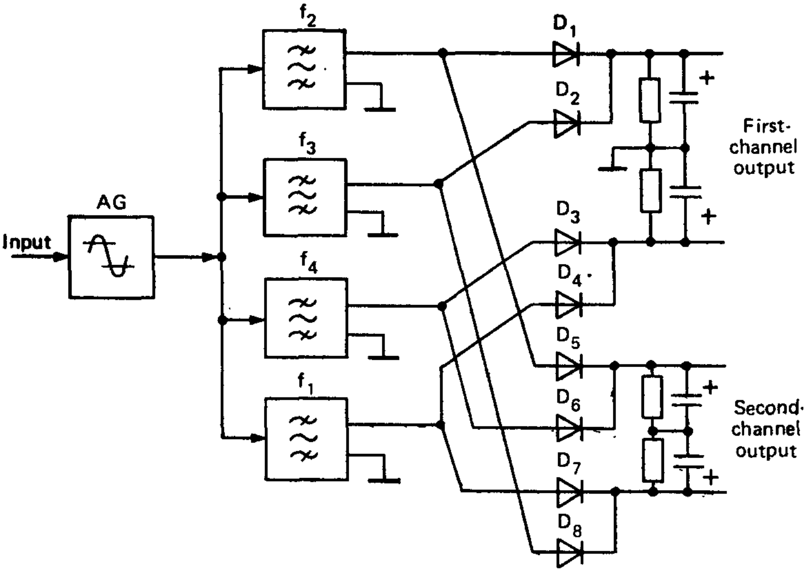


Fig. 9.17

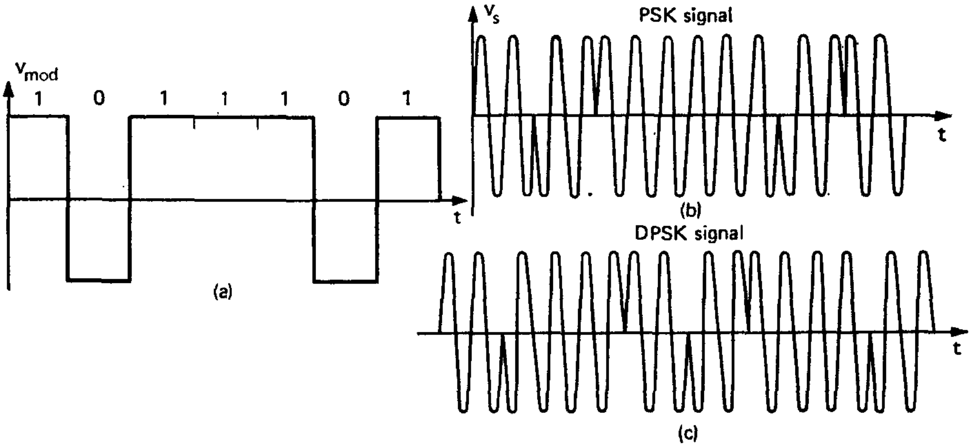
It uses four filters and eight diodes, D_1 through D_8, which are connected to the loads of the two channels in an appropriate manner. During the reception of a signal at frequency f_1, the current flows through diodes D_4 and D_7, producing across the loads a negative voltage drop which represents a binary 0 in both channels. During the reception of a signal at frequency f_4, the current flows through diodes D_3 and D_6, producing across the loads a negative voltage drop in the first channel (thus representing a binary 0), and a positive voltage drop in the second channel (thus representing a binary 1), and so on.

9.7. Reception of Phase-Shift-Keyed Signals

Phase-shift keying (PSK) is a form of phase modulation in which the modulating function shifts the instantaneous phase of the modulated carrier between two predetermined discrete values, $\Delta\varphi_1$ (when transmitting, say, a binary 1) and $\Delta\varphi_2$ (when transmitting a binary 0). Similarly to double-channel FSK described in Sec. 9.6, it is possible to use double-channel PSK in which case one uses four discrete phase-shift values: $\Delta\varphi_1$, $\Delta\varphi_2$, $\Delta\varphi_3$, and $\Delta\varphi_4$. As an example, consider single-channel PSK in which case the carrier phase undergoes a reversal, that is, a change of 180°, as shown in Fig. 9.18*a* and *b*:

$$\left.\begin{aligned} V_{s1}(t) &= V_{m,\,s}\cos(\omega_s t+\varphi_s)\\ V_{s2}(t) &= V_{m,\,s}\cos(\omega_s t+\varphi_s+\pi)\\ &= -V_{m,\,s}\cos(\omega_s t+\varphi_s),\quad 0\leqslant t\leqslant T \end{aligned}\right\} \qquad (9.18)$$

A PSK signal is detected by a phase detector arranged as shown in Fig. 9.19. It is fed the signal voltage defined by Eq. (9.18) and a



(a)

(b)

(c)

Fig. 9.18

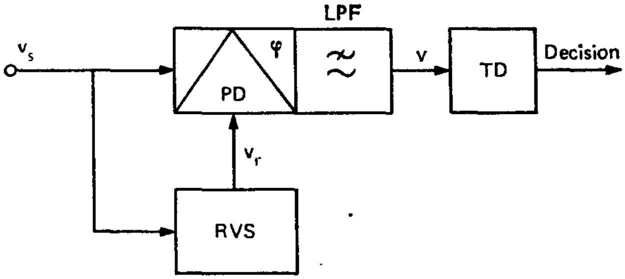


Fig. 9.19

reference voltage at the signal frequency and a constant phase such that

$$v_r = V_{mr}\cos(\omega_s t+\varphi_r)$$

It is supplied by a reference voltage source, *RVS*. The output of the phase detector consists of bipolar (plus-and-minus) d.c. pulses

$$v = \pm kV_{m,s}V_{m,r}\cos(\varphi_s-\varphi_r) \qquad (9.19)$$

The threshold device, *TD*, can operate at a threshold of zero, as in FSK. It is seen from Eq. (9.19) that the output voltage of the phase detector is proportional to the cosine of the phase difference

$$\Delta\varphi_r = \varphi_s - \varphi_r$$

and this imposes severe constraints on the performance of the reference voltage source. In contrast to FSK where a one-to-one correspondence exists between the transmitted symbol (1 or 0) and the information-bearing parameter (frequency f_2 or f_1), in PSK one has what may be called a reverted mode of operation. PSK signals lack a feature which could enable one to tie in the phase of the reference voltage with the signal phase. When the reference voltage is derived from the received information-bearing signal, its phase has two stable states, φ_r and $\varphi_r + \pi$. A change from one state to the other may be caused by an interfering signal and entail, in turn, a polarity reversal of the voltage at the output of the phase detector. To avoid this, resort is made to special coding schemes. In the case at hand, one may use what is known as *differential phase-shift keying*, or DPSK. In DPSK, the signal phase is reckoned from the phase of the preceding element. When a binary 0 (a "space") is transmitted, the signal phase remains the same as that of the preceding symbol, whereas in the transmission of a binary 1 (a "mark"), the signal phase is reversed, that is, changed by π, as shown in Fig. 9.18*c*. Each transmission begins by sending an element which bears no information and only serves as reference for phase comparison of the next signal.

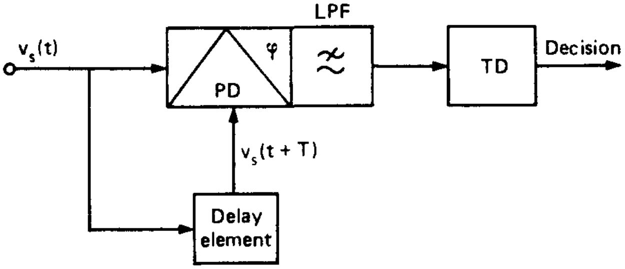


Fig. 9.20

DPSK signals can be detected in several ways. The most commonly used techniques are *phase-comparison* or *differentially coherent detection* and *polarity-comparison* or *coherent detection*.

In phase-comparison detection illustrated in Fig. 9.20, the reference signal is $v_s\ (t + T)$, that is, the input signal delayed for a time interval equal to the duration T of an elementary signal before it is applied to the phase detector. This completely eliminates any likelihood of 'reverted' operation.

In polarity-comparison detection, the signal is detected as in the case illustrated in Fig. 9.19, but the decision is made by a device arranged as shown in Fig. 9.21, by comparing the signal restored past the threshold device and the signal delayed by the time T for polarity. The polarity comparison is effected by a sign multiplier. This rules out the possibility of 'reverted' operation because each phase shift is accompanied by a polarity reversal of both voltages fed to the polarity comparator, *PC*. However, a phase jump is accompanied by an error. Worse still, each error in a DPSK system is doubled, and this impairs noise immunity.

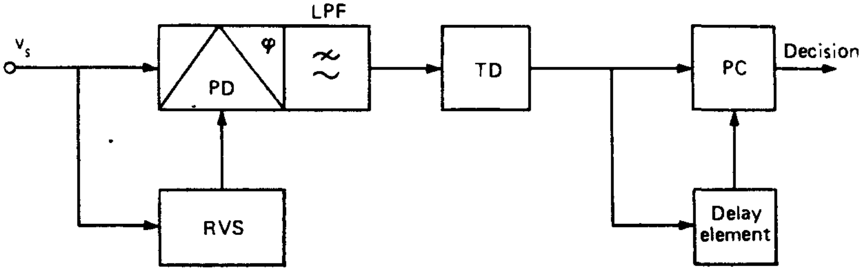


Fig. 9.21

An important subunit in a PSK receiver is the device that generates the reference voltage for the phase detector. Most often, the reference voltage is derived as the input-signal frequency is multiplied, filtered, and divided. Frequency multiplication is accompanied by phase multiplication. With a phase-shift keying of π, the signal frequency is doubled, and the signal phase becomes equal to 2π, which means that the phase-shift keying is cancelled. After a division by 2, the carrier wave remains unkeyed. In the more sophisticated cases of multiple phase-shift keying (MPSK), it is assumed that

$$\Delta\varphi_k = 2\pi k/n$$

and the frequency is then multiplied by n. In contrast to multiple FSK, a valuable property of multiple PSK is the fact that an increase in the number of levels is not accompanied by an increase in the bandwidth. Unfortunately, the resultant decrease in the modulation index $\Delta\varphi$ causes an impairment in noise immunity.

As has been shown by a number of investigators, PSK provides for a maximum ultimate noise immunity. It is widely used in state-of-the-art high-performance (notably, satellite) communication systems.

Chapter Ten

An Outline of Various Receiver Types

10.1. Signals Used in Radio Communication and Broadcasting

The configuration and layout of a receiver depend on its purpose, the frequency band in which it is to be used, and the class of signals it is to handle. All of these factors are interrelated. For example, multichannel point-to-point radio-relay and satellite links are microwave systems which use broadband *frequency-division-multiplex* (FDM) signals or TV signals frequency modulated in the channel. Use is also made of digital *time-division-multiplex* (TDM) transmission, and its field of use is gradually expanding. These systems operate by FSK or PSK, and they are ordinarily assigned fixed frequency bands so that their receivers need not be tunable.

Distress-signal transmission, paging, dispatching, intercom and other radio systems used by fixed and mobile services usually operate in the VHF band. They are allocated several fixed frequencies, so the receiver will have several preset frequencies which can be selected by rotation of a switch or by hitting a key. In their case, FM has been used most often until quite recently.

Long-haul radio links operating in the HF band ordinarily use SSB transmission with from one to four independently transmitted sidebands. SSB channels are used for analog telephone communication or, more frequently, for digital transmission. It is widely practised to use these channels for non-telephonic services, such as voice-frequency telegraphy in which case these sidebands are frequency-shift-keyed. Receivers for such applications are made more or less versatile. They can operate over a wide frequency range (say, from 3 to 30 MHz) and have end devices to process various signal types.

The versatility is still more pronounced in fixed and, in some cases, mobile sound broadcast receivers. Many of them are designed to receive monophonic and stereophonic programs, both AM and FM in the frequency range from LF to UHF. Their design poses problems related to the fact that they should be manufactured on a mass scale and at a relatively low cost. Now that standardized high-scale-integration IC modules, including microprocessors are readily available commercially, the tuners of such receivers are engineered along the same lines as their counterparts in professional equipments, that is, including frequency synthesizers and automatic controls.

10.2. Sound Broadcast Receivers

Broadcast receivers may have either built-in or external speakers. It is customary in radio-receiver practice to turn out packaged units capable of producing only the first portion of the functions of a receiver and delivering either r.f., i.f. or demodulated information to some other equipment. Such packaged units are known as *tuners*, and they are designed to drive a separate a.f. power amplifier or a packaged a.f. unit which may be used for sound-record reproduction.

Broadcast receivers widely differ in performance (notably, fidelity) which has a direct bearing on their prices. Those in the high-fidelity (hi-fi) class give the best performance and the utmost in control convenience owing to a more sophisticated design. Mass-produced receivers fall in the lowest-priced category, are simpler in construction, give a lower quality of performance, and lack some of the functional capabilities provided by their hi-fi counterparts, for example, they are usually built for monophonic rather than stereophonic reception.

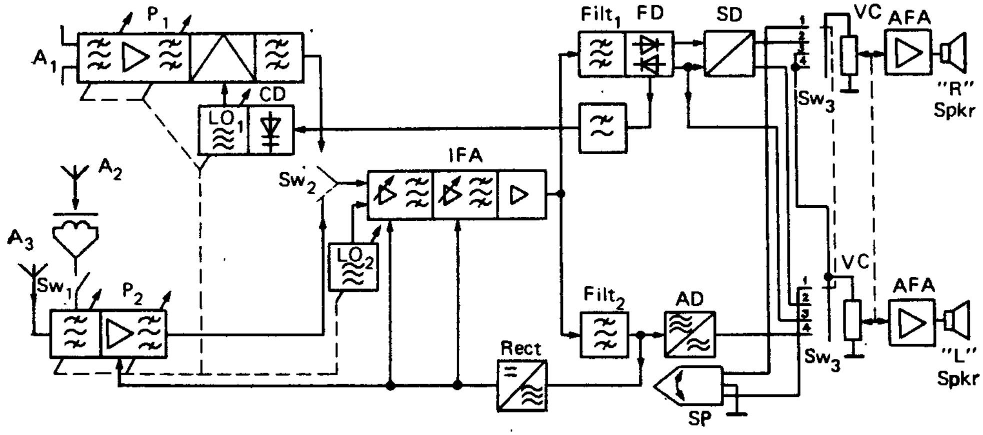


Fig. 10.1

A receiver capable of receiving AM programs in the LF, MF and HF bands and FM programs in the VHF and UHF bands must of necessity be built to have a two-channel configuration because AM and FM working obviously call for different operating characteristics. A simplified schematic of an AM-FM receiver is shown in Fig. 10.1. Here, P_1 is the preselector of the FM section complete with its antenna, A_1, and P_2 is the preselector of the AM section complete with its ferrite-rod antenna, A_2. Ordinarily, provision is made for connection of an external antenna, in which case the built-in antenna, A_2, is disconnected by a switch, Sw_1, which breaks as the plug of an external antenna, A_3, is inserted into the 'Ext' jack.

The i.f. amplifier may be made common to both sections. If this is the case, two series-connected tuned circuits are provided at the outputs of the amplifying devices instead of one i.f. tuned circuit or i.f. filter. One will then be tuned to the i.f. used by the FM section (quite often, this is 10.7 MHz) and the other to that of the AM section (ordinarily, 465 kHz). Such amplifier stages have a maximum gain at the two frequencies, which means that they are capable of amplifying both AM and FM signals. The passbands match the width of the signal spectra, both FM and AM. Choice of the signal to be fed from one of the two preselectors to the input of the i.f. amplifier is effected with the aid of a second switch, Sw_2.

The dashed lines in Fig. 10.1 show the ganged tuning arrangement. In this particular case, ganged tuning is effected continuously by a single tuning knob. Wide use is made of automatic and digital tuning (see Secs. 6.12 and 6.14).

In the reception of FM signals, the frequency converter uses local oscillator LO_1. In the reception of AM signals, the necessary frequency conversion is carried out by the first IFA stage in conjunction with a second local oscillator, LO_2 (the associated control circuits are omitted from the schematic).

The FM signal emerging from the i.f. signal is routed via the first filter, $Filt_1$, to a frequency discriminator, *FD*. Some of the voltage divides from the frequency discriminator to a low-pass filter and is applied to a control device, *CD*, to tune the first local oscillator automatically.

The AM signal is extracted by a second filter, $Filt_2$, and goes simultaneously to an amplitude detector, *AD*, and a rectifier, *Rect*. The rectified voltage is utilized for gain control. In the case at hand, the AGC loop includes the r.f. amplifier in the second preselector, P_2, and the early stages of the i.f. amplifier.

The output signal of the frequency discriminator goes to a stereo decoder, *SD*, which is used in the reception of stereophonic programs. The principles of stereophonic reception will be discussed a bit later.

The receiver has a mode-of-operation selector switch, Sw_3, a volume control, *VC*, and a two-channel a.f. amplifier, *AFA*, which drives two speakers, one in the left-hand channel, *LS*, and the other in the right-hand channel, *RS*. When the selector switch is in the '1' position, the a.f. section is fed signals from a stereophonic pickup, *SP*, when the a.f. section of the receiver is used in playing back stereophonic phonograph records. In the '2' position, the stereodecoder feeds signals to the a.f. amplifier which drives the left-hand and right-hand channels. In the '3' position, both a.f. channels are fed a monophonic FM signal directly from the output of the frequency discriminator. In the '4' position, the two a.f. channels are fed AM signals from the output of the amplitude detector.

10.3. Stereophonic Receivers

A stereo broadcast is produced by sending two signals designated as *L* (for left) and *R* (for right) from two (left and right) microphones spaced a suitable distance apart. At the point of reception, these two signals are reproduced by speakers also spaced some distance apart or by a stereophonic headset separately for the listener's left and right ears. Clearly, a stereo receiver must have two channels, left and right.

In the early stereophonic broadcast systems it was believed that a two-channel transmission needed an extended signal spectrum and, as a corollary, an extended bandwidth. That is why, this form of broadcasting has until quite recently been practised solely in the VHF and UHF bands which are well suited for high-fidelity sound transmission and reproduction and for noise-immune frequency modulation.

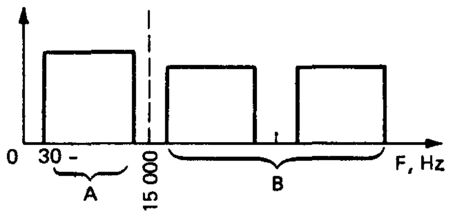


Fig. 10.2

State-of-the-art stereophonic broadcast systems are compatible in the sense that stereophonic programs can well be reproduced as monophonic by a monophonic receiver. In such a case, the modulating signal is derived from two spectra: a sum ($L + R$) signal transmitted using the complete audio bandwidth of up to 15 kHz, and an auxiliary signal transmitted on an ultrasonic subcarrier for the separation of the *L* and *R* channels. The spectral diagram of the multiplex signal thus produced is shown in Fig. 10.2. The ($L + R$) signal is transmitted within the *A* channel, whereas the *R* signal is fed to a balanced amplitude modulator which produces a suppressed-carrier AM signal at a frequency of 38 kHz (see Sec. 8.9). So that the suppressed carrier could be restored at the receiving end, a pilot signal whose frequency is 19 kHz (shown by the dashed line in Fig. 10.2) is additionally injected. At the receiving end, the pilot signal is separated by a filter and, after frequency-doubling, it synchronizes the reinserted 38-kHz carrier oscillator. The injection of the local carrier into the *B* spectrum turns it into an ordinary AM signal. After detection, it gives the '*R*' signal. For the stereo listener, the linear combination of ($L + R$) + ($L - R$) provides the *L* channel, and the linear combination ($L + R$) — ($L - R$) provides the *R* channel, thus completing the channel separation.

In the stereophonic system used in the Soviet Union, the modulating signal has the form shown in Fig. 10.3. The positive half-cycles are modulated by the *L* signal, and the negative half-cycles, by the —*R* signal. The mean value of the upper half-cycles of the signal varies in synchronism with the *L* channel, and that of the lower half-cycles in synchronism with the *R* channel. The difference of the two means gives a sum channel ($L + R$) and produces the *A* spectrum as shown in Fig. 10.2. The alternating component of the signal in Fig. 10.3 at the subcarrier frequency is modulated by both the *L* and the *R* signals and produces the *B* spectrum in Fig. 10.2.

The manner in which the two channels are separated is illustrated in Fig. 10.4. The modulating signal (see Fig. 10.3) is produced by a frequency discriminator, *FD*. In monophonic receivers, the speakers reproduce a sound which has the *A* spectrum. In stereophonic receivers, the output signal of the frequency discriminator, *FD*, is fed to a stereo decoder, *SD*, in which the positive and negative half-cycles are separated by diodes.

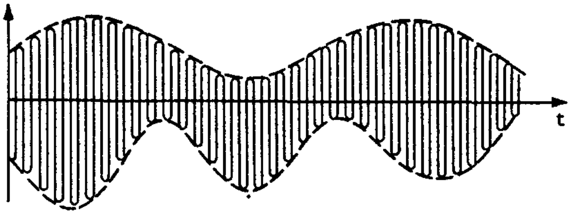


Fig. 10.3

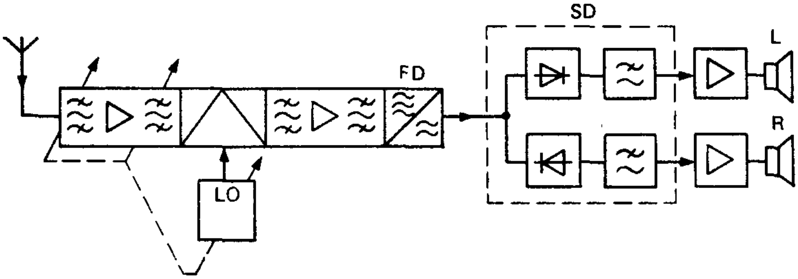


Fig. 10.4

The 1970s saw quite a number of attempts to transmit stereo programs in frequency bands hitherto used solely for AM transmissions, above all in the MF band. The same wide spectrum (up to 15 kHz) as is used in the VHF and UHF bands can hardly be transmitted at the frequencies allocated to broadcasting in this band. Still, the results obtained to date show that stereophonic transmission can be a success even in this band.

A major bottleneck in using the MF band for stereo broadcasting lies, as before, in the fact that two different signals have to be transmitted in the same common band of frequencies. Unfortunately, the two signals cannot be spaced far apart in frequency in the band as is done in the VHF and UHF bands (see Fig. 10.2). Since the width of the frequency channel is limited, the spectra of the two signals have to overlap. Also, the system must remain compatible, which means that the transmitted signal should normally be amplitude-modulated by the $L + R$ sum.

Several approaches have been tried in tackling the problem.

1. Both frequency and amplitude modulation is used. The signal is amplitude-modulated by the $L + R$ sum. The amplitude detector used in monophonic receivers does not respond to FM, and it delivers a complete monophonic signal. In a stereophonic receiver, the signal is additionally demodulated by a frequency discriminator, and the resultant voltage is frequency-modulated by the $L - R$ difference signal. The L and R signals are separated by adding together and subtracting the voltages from the two detectors.

2. Instead of frequency modulation, one uses phase modulation in combination with amplitude modulation. There is a phase detector to produce a difference $L - R$ signal. The reference voltage for phase detection is supplied by an automatically controlled local oscillator.

3. What is known as *quadrature modulation* is used. The transmitted signal consists of two AM signals whose carriers are at the same frequency but are shifted through 90 deg in phase relative to each other. Hence the name 'quadrature modulation'. One carrier is modulated by the L signal, and the other, by the R signal. The resultant voltage is thus both amplitude and phase modulated; which fact can readily be verified by plotting a phasor diagram for the multiplex signal.

As in the system in (1), the amplitude detector of a monophonic receiver produces a sum $L + R$ signal, which is what is required. In a stereophonic receiver, the multiplex signal divides into two channels, each of which contains a synchronous (phase) detector. The reference voltage fed to one of the detectors is 90 deg apart in phase from the voltage fed to the other detector.

As has been noted in Sec. 5.8 (see Fig. 5.1), when the reference voltage is in phase-quadrature with the detected signal, its output voltage is zero. That is why in the system in question one of the phase detectors responds to the signal that is in phase with its reference voltage and does not respond to the signal whose carrier is 90 deg apart in phase from the reference voltage. Conversely, the second detector responds to the second signal and blocks the first. As a result, the L and R signals are separated.

4. The multiplex signal is transmitted, using two independent sidebands and a common carrier. The transmission is arranged so that the L signal is carried by the lower sideband, and the R signal by the upper sideband. In a monophonic receiver, the detection of sucn a signal by a conventional amplitude detector produces a sum $L + R$ signal. In a stereophonic receiver, the multiplex signal divi-

des between two channels each of which contains a SSB detector. One channel receives the lower sideband and the carrier, and the other channel, the upper sideband and the carrier.

A feature common to all of the four systems examined above is that the multiplex signal is both amplitude and frequency modulated.

Consider two-channel transmission using both amplitude and angle modulation.

Suppose that there are two tone signals *1* and *2*, whose angular frequencies are Ω_1 and Ω_2. Then the signal voltage may be written as

$$v\,(t) = V\,(1 + m \cos \Omega_1 t) \cos\,[\,\omega t + \xi\,(t)]$$

such that

$$d\xi\,(t)/dt = \Delta\,\omega \cos \Omega_2 t$$

that is,

$$\xi\,(t) = (\Delta\omega/\Omega_2) \sin \Omega_2 t$$

When the angle modulation index is small, such that

$$\xi\,(t) = (\Delta\omega/\Omega_2) \ll 90^\circ$$

we have

$$\cos\,(\,\omega t + \xi) = \cos\ \omega t + \mu\ (d \cos\ \omega t\, dt) + \dots$$

$$\approx \cos\ \omega t + \xi\,(t) \sin\ \omega t$$

that is,

$$v\,(t) = V\,(1 + m \cos \Omega_1 t)\ \{\cos\ \omega t$$
$$-\ [(\Delta\omega/\Omega_2) \sin \Omega_2 t] \sin \omega t\}$$

It is seen that the radio signal spectrum still contains the carrier with an angular frequency ω and the sidebands which differ from the carrier by the values of the angular modulation frequencies, Ω_1 and Ω_2. As in conventional AM, the amplitudes of the components with frequencies $\omega \pm \Omega_1$ are half the peak value, that is, $V_m/2$. The relatively small quadrature component, sin ωt, of the signal is amplitude-modulated as

$$-V\,(1 + m \cos \Omega_1 t)\ [(\Delta\omega/\Omega_2) \sin \Omega_2 t] \sin \omega t$$

Owing to this component, the spectrum additionally acquires components with frequencies $\omega \pm \Omega_2$, whose amplitudes are $V\,(\Delta\omega/\Omega_2)$. Furthermore, the product

$$Vm\ (\Delta\omega/\Omega_2) \cos \Omega_1 t \sin \Omega_2 t \sin \omega t$$

is responsible for the presence of components with angular frequencies $\omega \pm (\Omega_1 \pm \Omega_2)$.

These components may lie beyond the receiver bandwidth if the sum $\Omega_1 + \Omega_2$ exceeds the maximum angular frequency, $\Omega_{\max}$, of the transmitted sound signal, lying at the edge of the bandwidth. However, these components are of the second order of smallness, and their suppression will not cause appreciable distortion in the stereophonic effect.

If an FM receiver has a frequency response which does not remain flat over the band, the FM may be converted to a spurious AM. As can be seen from Fig. 9.3, if the receiver is tuned exactly to the received-signal frequency, the AM frequency will be twice the FM frequency. After the amplitude detection has taken place, this spurious AM will show up as a distorted second-channel signal. Thus, the message can be transferred from the auxiliary to the main channel in which amplitude detection is effected.

As is seen from Fig. 8.3, with a receiver not exactly tuned to the received-signal frequency, the gain will be different for the left and right sideband frequencies of the AM signal spectrum, thus leading to a disbalance between the two sideband components. A signal having such an asymmetrical spectrum will change in both amplitude and phase. Thus, the amplitude modulation effected in the main channel can produce a response in the other channel effecting phase detection.

As follows from the foregoing, it is highly desirable for the predetector section of a receiver to have a flat frequency response. If this requirement is satisfied, the likely crosstalk and distortion will be negligible. Furthermore, the likelihood of crosstalk shows that double modulation ought not to be used for the transmission of two distinct messages within a common frequency band. For stereophonic transmission, however, such a technique is acceptable because the two signals carry the same intelligence, and the crosstalk can only slightly affect the stereophonic effect.

10.4. HF Broadcast Receivers

A typical block diagram of an HF broadcast receiver is shown in simplified form in Fig. 10.5. In this band, it is difficult to tune a receiver 'by ear', and it is still more difficult to keep the receiver 'on tune' if its local oscillator has a manual continuous tuning control. This explains why the required heterodyne (injection) frequencies are quite often produced by a digital frequency synthesizer in which the voltage-controlled oscillator (VCO) is servoed by a phase-locked loop (PLL). The voltage generated by the PLL is applied to the control circuit, *CS*, vis a low-pass filter, *LPF*, and a summator, Σ.

This is an infradyne receiver (see Sec. 4.5 and Fig. 4.6*b*) which fact simplifies the input circuit, *IC*, and it takes the form of a simple low-pass filter. Furthermore, this simplifies frequency synthesis and PLL operation because the VCO frequency may only vary within narrow limits. As has been noted in Sec. 4.5, in order to achieve this goal, it is customary to choose the first i.f. so that it lies well above the upper frequency limit of the receiver, say, of the order of 50 MHz. Accordingly, the first i.f. amplifier, IFA_1, is tuned to that frequency. It is followed by a second mixer, Mxr_2, and the second i.f. amplifier, IFA_2, which provides the major portion of signal amplification.

The local oscillator, *LO*, of the second frequency converter operates at a fixed frequency which differs from the first i.f. by the value of the second i.f. In the case at hand, the local oscillator is crystal-controlled.

The signal appearing at the output of the second i.f. amplifier is detected by an amplitude detector, *AD*. Quite often, this is a simple

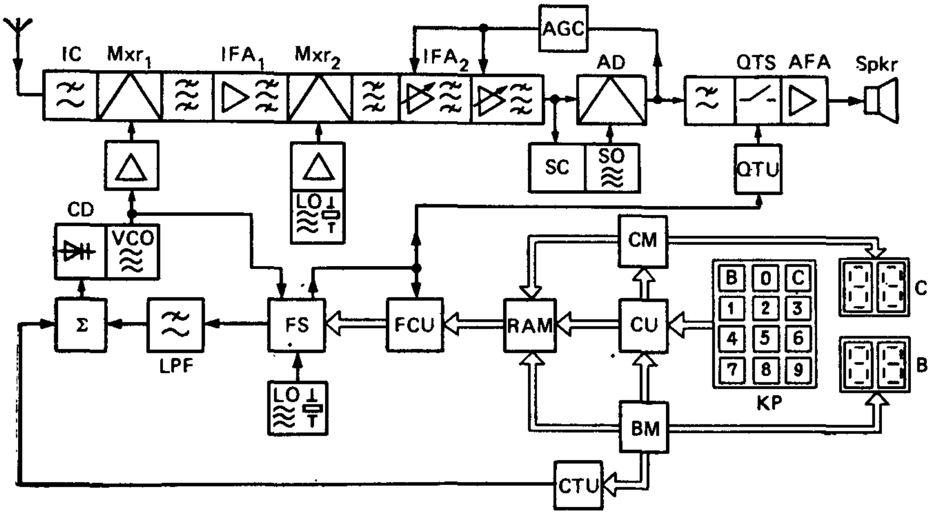


Fig. 10.5

diode detector, but the one shown in the diagram is a synchronous detector (see Sec. 8.7) complete with a synchronous oscillator, *SO*, and a synchronization circuit, *SC*. The latter separates the carrier from the received signal, and this carrier is then used by the PLL of the synchronous oscillator. The use of a synchronous oscillator serves a two-fold purpose. Firstly, this reduces the effect of interference and noise. Secondly, an opportunity is offered for the reception of SSB signals. True, SSB broadcasts in the HF band are not yet used, but this may quite well happen in the future.

The d.c. component of the detected signal is utilized in the AGC loop to generate the voltage used for gain control of the second i.f. amplifier. The a.c. component is fed to an a.f. amplifier, *AFA*.

When the receiver is being tuned from one station to another, no control voltage exists in the AGC loop, and the gain increases. This is accompanied by an increase in the noise so that it is heard on the speaker, *Spkr*, at about the same volume as the wanted signal would be heard in normal reception. To avoid this, it is quite often practised to provide a quiet-tuning circuit. In the case at hand, the circuit contains a quiet-tuning switch, *QTS*, operated by a quiet-tuning unit, *QTU*. While the receiver is being tuned, the switch is opened, and noise cannot reach the a.f. amplifier.

Digital tuning control is effected by changing the division ratio of the variable-ratio frequency divider in the frequency synthesizer, *FS*. In the HF band, radio broadcasting is assigned several bands divided into frequency channels spaced, as a rule, 5 kHz apart, and any of these channels can be used by a station. Each of these channels will usually be assigned a particular code number, and this number can be keyed in by operating a keypad, *KP* (as an alternative, this may be a soft-touch panel, see Sec. 6.17). As the keys of the keypad are pressed, a command is generated, which goes to a control unit, *CU*, and this in turn writes the selected band No. in a band memory, *BM*, and the selected channel in a channel memory, *CM*.

For band selection, the operator first presses the band select key, *B*. Following that, he hits the keys on the keypad, numbered from 0 to 9, in order to key in the desired band No. The number thus keyed in is transferred from the band memory unit to the band readout, *BR*. At the same time, these data are read into a RAM and to a coarse tuning generator, *CTG*. The latter generates an appropriate voltage which is routed via the summator, $\sum$, to the control device, *CD*, and sets the VCO to operate in the selected frequency band.

After the desired band has been displayed by the band readout, the operator presses the channel select key, *C*. As the next step, the operator keys in the desired frequency channel number which is then displayed by the channel readout, *CR*. Concurrently, the respective command goes via the control unit, *CU*, to the RAM. From the RAM, the tuning data are read into the final control unit, *FCU*, which controls the frequency synthesizer as appropriate. During the execution of the command, a signal is applied to the quiet-tuning unit, *QTU*, and opens the quiet-tuning switch, *QTS*. After the receiver has been tuned as described above, the switch closes, and the program from the station tuned in can be heard on the speaker.

10.5. Television Receivers

The TV signal bandwidth exceeds 6 MHz, therefore TV programs are transmitted in the VHF and UHF bands. The signal spectrum contains the video (or vision) and audio (or sound) signals, with the video signal reproduced as a picture on the screen of the picture tube or kinescope, while the sound signal is reproduced by a speaker. For these reasons, a TV receiver is of two-channel construction.

The video, or vision, signal is transmitted amplitude-modulated. As a way of saving the bandwidth, only one side-band is transmitted intact, and just a small fraction of the other sideband lying next to the carrier is transmitted. The sound signal frequency-modulates a subcarrier whose frequency is chosen to lie above the maximum upper frequency of the video spectrum. An approximate spectrum of the composite TV signal is shown in Fig. 10.6. Here, f_v is the vision carrier frequency, and f_s is the sound centre frequency. Colour TV systems are made compatible, which means that a typical, unaltered monochrome (or black-and-white) TV receiver can receive substantially normal monochrome from the transmitted colour TV signal.

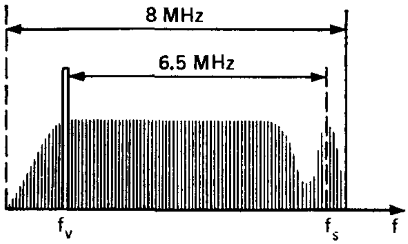


Fig. 10.6

Since the sound channel occupies only a small fraction of the composite signal bandwidth (see Fig. 10.6), the vision and sound signals are extracted from the composite signal, converted, and amplified together in the common pre-detector section of the TV receiver, and are separated in the output circuits. A simplified diagram of a typical monochrome TV receiver is shown in Fig. 10.7.

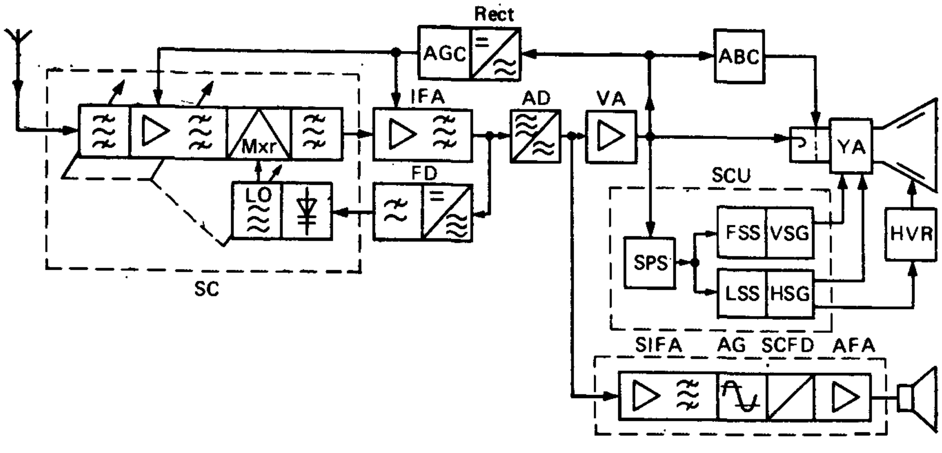


Fig. 10.7

The input unit, usually called the TV channel selector, *CS*, includes the input r.f. circuit and a frequency converter which in turn consists of a mixer, *Mxr*, and a local oscillator *LO*. The resonant circuits are tuned with the aid of varactors which are fed appropriate control voltages. The number of voltages is the same as the number of TV channels used. They can be generated by a digital electronic tuning unit or taken from potentiometers (not shown in the diagram).

The i.f. signal appearing at the output of the TV channel selector is then amplified by the i.f. amplifier and goes to an amplitude detector, *AD*, and, at the same time, to the frequency discriminator, *FD*, of the AFC loop controlling the local oscillator. The vision signal appearing at the output of the amplitude detector is amplified and applied to the cathode of the picture tube to control the intensity of the electron beam in proportion to the luminance of the picture elements. At the same time, the amplitude-detector output signal divides into the sound reproduction channel where it is amplified by the sound i.f. amplifier, *SIFA*, tuned to 6.5 MHz.

As is seen from Fig. 10.6, the sound centre frequency, f_s, is separated from the video carrier frequency f_v by 6.5 MHz. That is why the amplitude-detected signal contains both the vision signal spectrum and the sound-signal difference frequency, $f_s - f_v = 6.5$ MHz. It is this signal that is extracted by the sound i.f. amplifier. Thus, with regard to the sound signal the amplitude detector operates as a frequency converter for which the injection, or heterodyne, frequency is the vision carrier. The signal emerging from the sound i.f. amplifier goes to the sound-channel frequency discriminator, *SFD*, containing an amplitude limiter, *AL*. Finally, it is amplified by an audio-frequency amplifier, *AFA*, and reproduced by a speaker, *Spkr*.

The vision signal appearing at the output of the video amplifier, *VA*, divides into the sweep-circuit unit, *SCU*. Here, the sync pulse separator, *SPS*, separates the synchronization (or sync) pulses transmitted along with the sound and vision signals. As the next step, the composite sync signal is separated into horizontal (line) and vertical (frame) sync pulses by a line sync separator, *LSS*, and a frame sync separator, *FSS*, respectively. The line sync pulses control the horizontal (or line) scanning generator, *HSG*, and the frame sync pulses do the same for the vertical (frame) scanning generator, *VSG*. The currents from the two generators are fed to the yoke assembly, *YA*, and control the horizontal and vertical deflection of the electron beam. Additionally, the line scanning voltage is applied to an H.V. rectifier, *HVR*, from which the rectified voltage is applied to the anode of the picture tube.

Also, the output voltage of the video amplifier is fed to the AGC unit which contains another rectifier, *Rect*, and to the automatic brightness control unit, *ABCU*, from which the voltage is applied to the control electrode (grid) of the picture tube.

In colour television, the red (R), blue (B) and green (G) components of the televized scene are derived by means of three colour filters and separate camera tubes. The compatibility of colour television is ensured by transmitting a common luminance signal, L, as the sum of the three ($R + B + G$) signals, or the composite colour signal, which occupies the greater proportion of the allocated bandwidth (see Fig. 10.6). This signal is similar to that transmitted in monochrome TV and is normally reproduced by a monochrome TV receiver. In addition, the video spectrum includes subcarriers modulated by the colour-difference ($L - B$ and $R - L$) signals. The subcarriers, their amplitudes and other parameters are chosen so as not to affect the quality of the monochrome picture.

As in a monochrome TV receiver, the luminance signal is taken from the output of the video amplifier, *VA* (see Fig. 10.7), to the cathodes of the picture tube (which is in this case a three-beam colour type) and controls the brightness of the picture elements on the screen. At the same time, this signal is fed to what is called a colour-identification unit not provided in a monochrome receiver. In this unit, the colour-difference signals are separated and combined linearly so as to produce signals of the form '$R - L$', '$G - L$' and '$B - L$', which are applied to the control electrodes (grids) in the three electron guns of the colour picture tube. As a result, the signal that exists between the cathode and grid in one electron gun is $L + R - L = R$, whereas in the second the net signal is B, and in the third, G, according to the scene being televized, and provides for faithful reproduction. Furthermore, a colour TV receiver contains a unit which causes the three electron beams of the three-gun colour picture tube to converge to picture elements on the screen. Quite aptly, it is called the beam-convergence control.

The configuration of a colour TV receiver from the antenna to the detector in the vision channel and as far as the speaker in the sound channel remains basically the same as that of a monochrome TV receiver (see Fig. 10.7).

10.6. Long-Distance Communication Receiving Stations

The line, satellite and radio-relay links currently in operation have between them a capacity of many thousands of telephone channels and can handle large flows of digital data and TV programs. All in all, they are quite capable of meeting the need for long-distance communication. Still, point-to-point HF radio links remain an important element of radio communications.

Radio facilities operating in this frequency band provide communication with remote mobile objects, small communities, and hard-to-reach localities. They are also useful as a back-up means because they can transmit and receive information directly over any distance, however long. Owing to the reflection of radio waves from the ionosphere, the range of radio communication in the HF band may be as great as 10 thousand kilometres or even more with only one midway relay station or without such a station at all.

Long-distance radio links handle information mostly in coded form. The radio telephone service in the HF band is of limited use because it lacks privacy, unless sophisticated measures are taken to make unauthorized reception difficult. It is no less important that telephone signals need a large bandwidth. Given the same bandwidth, telegraphy can handle a much greater traffic.

As a rule, point-to-point receivers and ancillary plant are located tens of kilometres away from point-to-point transmitters. This is because the HF point-to-point transmitters radiate large blocks of power and, if they were located close to point-to-point receivers, this would cause severe interference due to nonlinear effects (crosstalk and intermodulation).

Ordinarily, a receiving radio station is situated outside the city limits and away from sources of interference. The station has service buildings, an antenna array, and a power supply system. The antennas are of the highly directional type to avoid or at least attenuate interfering signals. As a rule, one and the same station is designed to provide communication in different directions. Also, for diversity reception (see Sec. 7.12) in one direction there should be two or more antennas. That is why the total number of antennas in the antenna array may run into several tens. The down-leads from the antennas to the station buildings are protected by lightning arrestors.

The antenna of each direction can be used for communication over a wide range of frequencies. If a receiving antenna were connected to only one receiver, it would be necessary to provide individual antennas for other transmissions in the same direction. Because of their complexity and size, this would cost more and call for an extended antenna array and a greater distance from the antennas to the service building, thus leading to a greater loss of signal power in the connecting lines. To avoid all or at least part of this, each antenna is designed to serve several radio links.

If an antenna were connected to several receivers directly, the signal power at the input to each would be reduced. Also, it would be difficult to match the connecting lines (feeders or cables) to the receivers, might cause interference between the input circuits of the receivers and crosstalk due to the local-oscillator signals from the receivers. Instead, the line from each antenna is connected to an amplifier which has separate outputs for the various receivers.

An antenna amplifier is to meet a number of specific requirements which may be summed up as follows.

* Its bandwidth should be just large enough to encompass all the frequency range within which a given antenna can receive signals.

If its bandwidth is made too wide, it might include interfering signals from high-power transmitters.

* Its electron devices ought not to display their nonlinearity.
* There should be no cross-interference between the receivers connected to the same antenna amplifier.

The way in which the antennas of a receiving radio station can be connected and switched is shown in Fig. 10.8. Here, A_1 through A_n are the antennas. The feeder lines, *FL*, from the antennas are carried to broadband antenna amplifiers, *BAA*. Each amplifier has several independent outputs connected to the buses of an antenna commutator, *AC*. The vertical buses of the commutator are connected to receivers Rvr_1 through Rvr_m. The commutator can connect any receiver to any antenna. Every measure is usually taken to minimize power losses at the contacts, because otherwise there would be an increase in the noise factor of the receiving system.

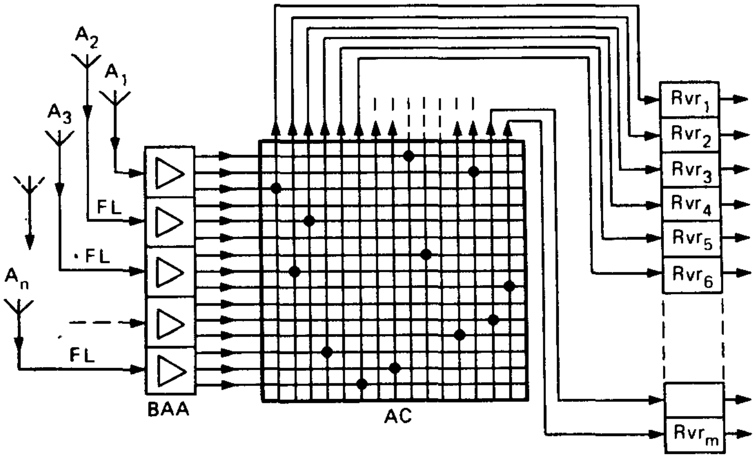


Fig. 10.8

The output signals of the receivers are transferred over a cable or radio-relay link to a processing centre which also distributes the signals among the end users and monitors the quality of communication. The signals can be relayed to the recipients via a central telegraph office or an intercity telephone exchange. There is also equipment for multichannel transmission of the received signals.

Now that HF radio stations haye to carry only a small fraction of the overall point-to-point radio traffic, it is becoming increasingly more important to cut the operating costs. This can be done by reducing their staff. This trend is taken into account in the design of state-of-the-art receivers, antenna commutators and other pieces of equipment; it is designed for remote or programmable automatic control and unattended operation. This approach improves control flexibility and service reliability.

10.7. Point-to-Point Radio Receivers

Long-distance radio communication in the HF band came into being more than a half-century ago as a multifunctional facility for information exchange. To a marked extent, it still retains this function even today. That is why point-to-point receivers are usually designed to handle several kinds of signals, the predominant types being digital, coded, and telegraph signals. Their transmission uses amplitude-shift keying (ASK), frequency-shift keying (FSK), and phase-shift keying (PSK), but provision is also made for the reception of AM and SSB telephone signals.

In SSB operation, one radio link can use from one to four independent sidebands, depending on the conditions of radio wave propagation and the presence and level of interference. When three or four SSB channels are used, the channel width is 3 kHz, which is quite enough for commercial telephony. More often, however, SSB channels are simultaneously used for telephone and non-telephonic services, such as voice-frequency telegraphy or facsimile transmissions on a subcarrier modulated by the picture signal. To be able to process such different signals, multifunction receiving stations include appropriate output devices. Now that radio links operating in the HF band have come to serve a limited range of functions owing to advances in the other bands and other systems, it is likely that fewer signal types will have to be handled. This is bound to lead to simplifications in receiver design and performance optimization.

Point-to-point receivers come in several standard layouts. When describing and comparing them, it is usual to divide such a receiver into the following basic subunits:

— the main reception section;

— a device intended to supply heterodyne (injection) voltages and to provide the tuning function;

— an end (or output) device or devices;

— receiver controls;

— a device intended to monitor the signal quality and the operating conditions, and to detect any faults in the key subunits and assemblies of the receiver;

— regulated (stabilized) power supplies.

For two-branch diversity reception, it is usual to place two receivers in a common rack or enclosure and to add a signal combiner.

The main reception section includes an r.f. amplifier, a frequency converter (or converters), i.f. amplifiers, and detectors. The input of the main reception section is connected via an antenna commutator (see Fig. 10.8) to one of the outputs of a broadband antenna amplifier, and the output signals of the detectors are fed to an end device (or devices) for further processing as appropriate.

The device intended to supply injection (heterodyne) voltages and to effect the tuning function contains, above all, a controlled frequency synthesizer. It may also include sources of control voltages for electronic control of tunable or adjustable frequency-selective circuits in the receiver. This job may, for example, be done, by the input pre-selection filters.

In the end devices, the multichannel group signals coming from the detector outputs are processed before they are conveyed to the users. This processing consists in that the messages that make up the multiplex signal are separated, the characters mutilated in the course of transmission are identified and corrected or replaced by correct ones, and the signals are converted to a format acceptable for the user.

In many cases, the main reception section is arranged as an infradyne, which fact simplifies the use of a frequency synthesizer and digital tuning. As has been noted in Chap. 4, one of the features of the infradyne is an untunable broadband input circuit. In fact, the passband of the r.f. circuits in an infradyne is so wide that it can pick up signals from many high-power transmitters, with a total voltage of several hundred millivolts. This might be responsible for interference in the early stages of the receiver owing to crosstalk and intermodulation. In high-sensitivity point-to-point receivere these phenomena may impair the quality of reception. To mitigats them, it is practised to place at the input several switch-selectable, but untunable bandpass filters which pass only that part of the spectrum in which the reception is going on at the moment. This brings down the interference level at the input and mitigates nonlinear processes.

Nonlinear interference may further be reduced as follows.

1. The early stages having a large bandwidth are arranged to have a low gain. As a result, the received-signal voltage is amplified there only as much as is necessary for the noise factor not to increase appreciably. Otherwise, there would be an increase in the effect produced by nonlinear processes because they are proportional to the noise voltage cubed (see Secs. 8.4 and 8.5).

2. An attenuator is placed at the input to the receiver. The point is that to make radio communication more reliable and stable, it is usual to operate transmitters with an ample power reserve. Because of this, they produce at the receiving end a strong signal substantially exceeding the sensitivity threshold of the receiver. If the interference level is very high, the attenuator reduces the incoming signal voltage to a level at which normal reception is still possible.

The interference level is brought down as well many times because the effect of the interference is proportional to the voltage cubed. The attenuator is controlled by an AGC loop.

10.8. Low-Noise Receivers

Radio-relay, satellite and space radio systems operate in the UHF and SHF bands. As a rule, the receivers used in line-of-sight multichannel radio-relay links are of the superhet type. From receivers operated at the lower frequencies, they mainly differ in the construction of the r.f. circuits which use waveguides, coaxial and strip lines and filters and have a large bandwidth and a high intermediate frequency (tens of megahertz).

At each terminal station of a radio-relay link, the received signal is detected and goes to a terminal equipment where it is separated on a frequency or a time basis, depending on whether the radio link uses frequency-division or time-division multiplexing in transmission. At repeaters where channels are neither inserted or dropped, the incoming signal is usually not detected. Rather, the r.f. channels are converted back to the i.f. band for amplification, then each is translated up and re-transmitted in a microwave frequency band different from the received frequency towards the next repeater, and so on.

A distinction of the microwave bands in which radio-relay links operate is the low level of external noise. Therefore, it is important for receiver noise to be as low as possible. If this requirement is satisfied, the resultant increase in sensitivity makes it possible to use transmitters of lower power ratings and of simpler design.

The sensitivity requirements are especially stringent for receivers used in satellite and deep-space radio communication systems. These requirements are met through the use of specially designed input amplifiers with a minimal noise temperature and a threshold-reducing devices for FM detection (see Chap. 9). Where an especially high sensitivity is essential, resort is made to cooled parametric amplifiers, such as shown in Fig. 4.33.

10.9. General Trends in Receiver Automation and Optimization

The advent of the transistors in the 1950s and 60s had brought with them deep changes in radio reception practices and equipment. A similar impact was produced in the 1970s by integrated-circuit (IC) technology. At first, integrated circuits, were standard combinations of about 100 components on a chip—a level known as *small-scale integration* (SSI). Then followed what is known as *medium-scale integration* (MSI), defined to have more than a hundred but less than a thousand components per chip, *large-scale integration*

(LSI) with over a thousand components per chip, and, finally, *very large-scale integration* (VLSI) with over ten thousand elements on a single chip. In the 1980s, chips with over one million components, referred to as ELSI (for '*extremely large-scale integration*') became a reality of life. One of the important consequences of IC technology has been the development of automatic control subunits based on digital techniques, microprocessors and microcomputers for receivers.

Another important "milestone" in electronics has been metal-oxide-semiconductor (MOS) technology which has offerred a tool for producing a great number of logic gates per unit area of a chip, a feature of primary importance for microprocessors largely made up of memory registers.

Microcomputers are built into receivers to form programmable-logic functional subunits. As an alternative, a microcomputer may be common to a bank of receivers at a receiving complex. At this writing, the capabilities of microprocessors and microcomputers as far as radio reception is concerned have not yet been fully explored and appreciated. It is beyond any shadow of doubt, however, that they are bound to bring about sweeping changes in receiver design and operation as well as in radio engineering as a whole. It is especially advantageous to use a microprocessor in cases where the task at hand involves arithmetic and logic operations, and a need exists for flexible control, statistical data processing, and similar functions.

Microprocessors may be used as stand-alones to perform relatively simple functions, or as the building blocks of more elaborate data-processing devices, or as devices that provide a tie-in between the functional elements of various complexes at a receiving radio station.

At the lowest level of this hierarchy, microprocessors are used to switch the input attenuators and other elements of r.f. circuits, to effect electronic control of variable-ratio frequency dividers in frequency synthesizers, and to assist in local and remote tuning control (including programmable tuning).

At the second level of the hierarchy, microprocessors are used to analyse the channel status, to assess the stochastic, power and descriptive parameters of the signal and disturbances for purposes of adaptive receiver control, to implement optimal signal-processing algorithms under conditions of *apriori* uncertainty, to build digital filters for signals of complex waveform and adaptive code converters in channels with isolated and group errors, to carry out an optimum synthesis of receivers in a multidimensional space of performance criteria, and in computer-aided design of receivers.

Microprocessors at the third level are used for mathematical modelling and computer-aided evaluation of equipment efficiency, intersystem interference compensation, channel and message switching, automated frequency allocation, automated data gathering from a great number of remote receivers, and as building blocks for data concentrators, intelligent terminals for radio operators, and some other purposes.

Microprocessors offer a means with which radio receivers can ultimately be optimized in terms of the key performance criteria. Let us take a closer look at the matter.

The concept of vector, or multi-objective, optimization with reference to a radio receiver has been outlined in Sec. 1.1. A device, D, treated as a system, will be referred to as an admissible one if it satisfies a set of functional constraints, C_f, and a set of structural and parameter constraints, C_s. The set of admissible systems, S_a, contains a subset, S_{sa}, of strictly admissible systems which satisfy an overall criterion (C_f, C_s, K, C_k), where K is the system quality (or performance) criteria vector, and C_k is the constraints on the quality criteria. The goal of vector or multi-objective optimization is to select from the subset S_{sa} a system, D, which has the best vector K in the sense of some criterion. The system $D \in S_{sa}$ is called the worst unless there is one unconditionally better system, $D_b \in S_{sa}$, for which K $(D_b) \leqslant$ K (D), that is, each of the individual quality criteria, k_i (D_b), is not worse (not greater) than k_i (D). If S_{sa} does not contain any unconditionally better system, the given system D is referred to as a better-than-worst one. By applying the unconditional preference criterion, UPC, to all $D \in S_{sa}$, it is possible to divide it into two non-interesting sub-sets: S_w, which contains the worst systems, and S_{btw}, which contains better-than-worst ones. Since the optimization procedure has as its goal to find better-than-worst systems, the subset S_w may be excluded from consideration.

In describing better-than-worst systems on the basis of the UPC, resort is made to the functional characteristic method and the weight method. Both methods reduce vector optimization to scalar optimization, thus permitting one to discard better-than-worst systems and to find individual potential values k_{i0} of each quality (performance) criterion. However, it is only in the case of a degenerate subset S_{bw} containing a single system, that it is possible to find the best system. Therefore, in order to choose the only system from the non-degenerate subset S_{bw}, it is practised to introduce a conditional preference criterion, CPC, at the final stage of the synthesis procedure. The most commonly used methods of vector optimization based on the CPC are as follows: the resultant quality criterion method, the Minimax method, the constraints method, the consecutive tradeoff method, and the expert estimate method.

A drawback of the resultant quality criterion method, which reduces the optimization problem to a scalar one, lies in the subjective approach in establishing the functional dependence of the overall resultant quality criterion on the individual criteria, $K_{res} = f(k_i, \ldots, k_m)$. Sometimes, the form of this relation cannot be substantiated even subjectively. In situations like that, one may resort to the Minimax method by which one can choose a system $D_M \in D_{sa}$ such that

$$k_{m,s}(D_M) \leqslant k_{m,s}(D)$$

where

$$k_{m,s} = \max(k_{1,s}, \ldots, k_{m,s})$$

is the largest of the specified quality criteria

$$k_{i,s} = k_i/k_{i,\max}$$

However, this only gives the best (least) value for the worst (largest) quality criterion.

The constraints method is based on the fact that all the individual quality criteria, except one called the principal criterion, are expressed as constraints in the form of equalities and inequalities, such as $k_2 = k_{20}, \ldots, k_n = k_{n0}$, or $k_2 \leqslant k_{2m}, \ldots, k_n \leqslant k_{n,m}$. With a sufficiently great number of constraints, the vector synthesis reduces to a scalar one, but the method itself involves a good deal of arbitrariness. Also, constraints of the equality type may lead to a solution belonging to the subset S_W.

In the consecutive tradeoff method, one ranks the quality criteria in the order of significance, finds the value $k_{1,\min}$ while ignoring all the other criteria, then chooses a tradeoff, that is, the permissible increase in k_1 and finds $k_{2,\min}$, and so on. At the final stage, one finds a system giving $k_{n,\min}$, subject to the constraints

$$k_i \leqslant k_{i,\min} + \Delta k_{i,\min}$$

on the remaining $n - 1$ criteria. In this case, the constraints imposed on the $n - 1$ minor criteria can be chosen in a more well-founded way, but here, too, there is some subjectivity in the choice of the principal quality criterion and of the tradeoff Δk_i.

The expert estimate method offers a means by which one can specify the initial data (C_f, C_s, K, C_k), define quantitatively C_s and C_k, choose the weight coefficients, etc. Yet, this method suffers from a number of drawbacks as well, related to the recruitment of experts and the organization of the estimation procedure itself.

Consider several examples of structural and parametric optimization.

Example 10.1. The decision circuit which has an important bearing on the quality of reception can be synthesized optimally if one knows the signal-to-noise ratio at its input or, which is the same, at the output of the linear section, or the main reception section (see Sec. 10.7). The quality of reception improves with an increase in this ratio. Therefore, it is of prime importance to optimize the decision circuit in the sense of the signal-to-noise ratio. This criterion is invariant to a certain degree towards the type of signals and how they are processed, and is related to other criteria of message reception quality, such as the probability of error and the mean-square value of distortion. After a certain formalization, this vector criterion makes it possible to take into account external conditions (transmitter power output, antenna directivity and orientation, radio wave propagation, interference and noise characteristics, etc.), the performance of the receiver proper (configuration, nonlinear distortion of the signal in the selective circuits, the relationship between the nonlinear distortion factor k_{nl}, the blocking interference factor k_{b}, the crosstalk interference factor k_{c}, the intermodulation interference factor k_{int}, the adjacent-channel and spurious response factor k_s, the parameters $\alpha_1, \ldots, \alpha_n$ of the decision circuit, the nonlinearity factor ξ_{nl} of the electron devices, antenna noise, receiver noise), and the constraints on the receiver structure and performance.

Most often, the primary criterion chosen to describe the receiver performance vectorially is the signal-to-noise ratio, h_{s}^2, at the output of the main reception section. To evaluate how perfect a real receiver is as regards its selectivity, it is customary to introduce the concept of an ideal circuit which possesses the selectivity of an optimum matched filter. Then the quality of reception is stated in terms of the error accumulation factor

$$k_{\text{err}} = h_{\text{s,ideal}}^2 / h_{\text{s,real}}^2$$

where the subscripts 'ideal' and 'real' refer to the ideal and real circuits. For discrete signals,

$$k_{\text{err}} = p_{\text{err,real}} / p_{\text{err,ideal}}$$

where p_{err} is the error probability. Analysis shows that

$$k_{\text{err}} = \{[1 + (N - 1)/\alpha_{\text{n,A}}] + h_{\text{s,ideal}}^2 (k_{\text{c}}^2 + k_{\text{int}}^2 + k_{\text{s}}^2)\} / (1 + k_{\text{nl}} + k_{\text{b}})^2 \quad (10.1)$$

where N = noise factor of the receiver

$\alpha_{\text{n,A}}$ = factor taking care of the real antenna noise level

If k_{err} is close to unity in value, this is an indication of the high quality of the main reception section. If, on the other hand, $k_{\text{err}} \gg 1$, the quality of reception can substantially be enhanced by improving the main reception section. Given a certain level of interference and noise in a radio link, preference should be given to a receiver having the lowest k_{err}. Equation (10.1) suggests the ways and means by which the receiver can be optimized as regards its structure and parameters. Since the quantities entering Eq. (10.1) are interrelated in a complex manner, the optimization procedure should be carried out on a computer.

Example 10.2. The quality of a point-to-point receiver can be evaluated on the basis of an integral estimate which takes into account the individual performance indices of the receiver and the external devices (the antenna, the power divider, the broadband antenna amplifier, etc.). For this purpose, one introduces the reception loss factor, $K_{\mathrm{r.l.}}$. The best receiver configuration is chosen on the basis of $K_{\mathrm{r.l.}}$ subject to the receiver cost C_{r}. Thus, one in effect uses a vectorial quality criterion, K ($K_{\mathrm{r.l.}}$, C_{r}); given the same $K_{\mathrm{r.l.}}$, the best receiver will be the one which has the lower value of C_{r}. In evaluating $K_{\mathrm{r.l.}}$, it is assumed that the HF band accommodates close on 9000 frequency channels 3 kHz wide each. It is also assumed that any of these channels is unusable for communication if the power of any type of noise (atmospheric noise, man-made noise, concentrated noise, receiver noise, etc.) at the channel output is not less than the design power of the signal whose frequency is the same as the allocated frequency. If interference from other stations is nonexistent and the input signal level in each of the N frequency channels is the same as the mean atmospheric noise level, then an average of 50% of the channels are unusable for communication because the noise would produce false signals at the output even in the absence of wanted signals. Therefore, in evaluating the quality of reception, one should proceed from the assumption that in actual conditions the input signal level is such that 50% of the channels is occupied by interference and noise. On the other hand, the HF band is heavily contaminated by unintentional noise and interference and the r.f. section has a nonlinear frequency response. Therefore, the input signal level has to be raised, if the same 50% of the channels are to be rendered usable for communication. The difference between the two levels, averaged over all the channels, is what we have called the reception loss factor, $K_{\mathrm{r.l.}}$. It is expressed in decibels referred to one microvolt, or dBμV. As an illustration, Table 10.1 lists the values of $K_{\mathrm{r.l.}}$ and C_{r} for several typical point-to-point receivers.

Table 10.1. **Performance Criteria of Point-to-Point Receivers**

Quality criterion	Receiver type						
	1	2	3	4	5	6	7
$K_{\mathrm{r.l.}}$, dBμV	1.6	13.1	10.5	9.8	10.5	12.7	10.3
C_{r} (relative units)	9	11	3.8	9.3	2.5	1.8	0.55

As is seen, the reception loss factor has an average value of $K_{\mathrm{r.l.}} \approx 11$ dBμV. This implies that for the above receivers the input signal level should be about 10 times than it should be for a receiver exposed solely to atmospheric noise. Note the drastic difference between the quality criteria, K (10.3, 0.55) and K (9.8, 9.3) for receiver types *7* and *4*, respectively: in terms of reception loss they differ by a mere 5%, whereas in terms of cost, they differ by a factor of 17. This indicates that the technical approaches used to improve receiver type *4* give an insignificant improvement in its performance but add very much to its cost. Obviously, an important goal in receiver optimization is to reduce the cost while reducing or maintaining constant the reception loss factor.

Conclusion

The ten chapters you have read have set forth only a general outline of radio receivers and of their further development. Radio reception is a most challenging task in radio engineering, and far from all approaches to its achievement have been fully explored and put to use. This is a gratifying field for the radio engineer to apply his creative power. Novel ideas and engineering approaches are reported almost every day. If we extrapolate the advances in radio reception in the past decades, it may well be expected that these new contributions are bound to lead to further changes in radio-receiving equipment in fifteen to twenty years from now, primarily due to advances in components. However, the bank of basic ideas and principles underlying receiver design undergoes far slower changes.

In the nearest future, as has been the case so far, solutions to many problems will depend on the development and commercialization of novel materials (primarily, semiconductors) and manufacturing techniques, and the wider use of computer-aided design procedures in the synthesis of sophisticated equipment.

One of the controversies still awaiting its resolution is the choice between analog and digital devices. On the one hand, digital pulse devices are a most perfect match for IC technology and, on top of that, possess the fullest capability to restore noise-multilated signals to their correct shape.

On the other hand, the 'digital' approach is not always the best one from an economic point of view. An apt example is the diode amplitude detector. Its analog variant consists of just three components: a diode, a resistor, and a capacitor. Quite a number of digital amplitude detectors have been developed to date, but their component count runs, as a rule, into at least several hundred, and the cost is higher in proportion. Another limitation is the fact that digital pulse-code-modulated (PCM) signals need a greater bandwidth. Also, the use of a great number of circuits carrying pulse currents which have a wide spectrum may lead to an increased receiver noise level.

In all probability, it will be most advantageous to combine analog and digital techniques, but their respective contributions and specific applications have not yet been completely explored and evaluated.

Nor should we rule out further breakthroughs in the field of radio receiver components, which are difficult to predict. They are quite likely to occur, for example, in IC technology. Much as the addition of another bit to a binary code alphabet doubles the number of likely code combinations or words, so the addition of another physical process opens up a broad field for the synthesis of components having novel important functional properties. From this point of view, almost inexhaustible resources exist for the synthesis of, say, opto-acousto-electronic components. For such resources to be utilized to advantage, the designer should have a working command of the bank of ideas accumulated by his predecessors and should clearly realize the scientific, technological and social goals and objectives in his specific field.

If the student is to be successful in mastering the fundamental physical principles of radio reception that have accumulated in the past decades, in applying the general methodology of radio receiver synthesis, and in implementing them in state-of-the-art circuit configurations, he must be well taught in the general sciences and trained in their practical applications. All of this is taken care of by college curricula, but it is vitally essential for the student to undertake research and development work on his own. The course in radio receivers offers broad possibilities for such an activity. Its subject-matter includes the modelling and analysis of processes varying in complexity: from the response of elementary linear circuits in the case of a simple signal to the extraction and processing complex weak signals in complex nonlinear electron devices in the presence of a mixture of various interfering signals and noise. Accordingly, there is a practically unlimited possibility for the student to select and solve problems and to obtain new results which may be of value to radio engineering as well as to the student, as a beginning researcher, himself. In his work, the student will surely get a good deal of help from the computer which has now become a common feature at colleges.

The satisfaction he gets from his own project, though modest but carried out on his own, and from a successful improvement or invention will fortify the young engineer's faith in his capabilities and creativity.

Index

TO THE READER

Mir Publishers welcome your comments on the content, translation, and design of the book.

We would also be pleased to receive any suggestions you care to make about our future publications.

Our address is:

USSR, 129820, Moscow, I-110, GSP, Pervy Rizhsky Pereulok, 2, Mir Publishers.

About the Publishers

Mir Publishers of Moscow publishes Soviet scientific and technical literature in many languages comprising those most widely used. Titles include textbooks for higher technical and vocational schools, literature on the natural sciences and medicine, popular science and science fiction. The contributors to Mir Publishers are leading Soviet scientists and engineers from all fields of science and technology. Skilled staff provide a high standard of translation from the original Russian. Many of the titles already issued by Mir Publishers have been adopted as textbooks and manuals at educational institutions in India, France, Cuba, Syria, Brazil, and many other countries. Books from Mir Publishers can be purchased or ordered through booksellers in your country dealing with V/O "Mezhdunarodnaya Kniga", the authorized exporter.